Journal of Modern Applied Statistical Methods

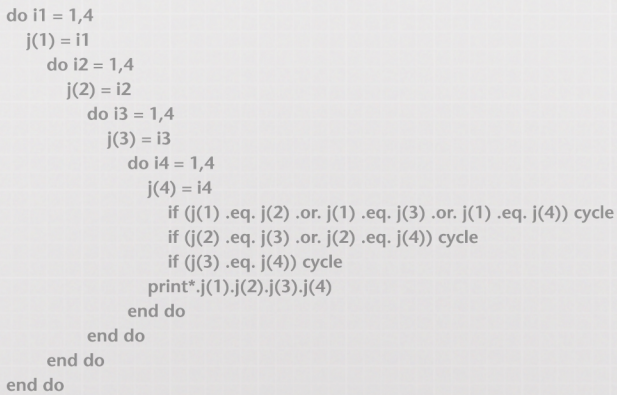
Volume 12 | Issue 2 Article 30

11-1-2013

Vol. 12, No. 2 (Full Issue)

JMASM Editors

בס"ד



Journal of Modern Applied Statistical Methods

Invited Articles
Florence Clark, H. J. Keselman,
Abdul R. Othman *and* Rand R. Wilcox

Vol. 12, No. 2 • November, 2013

בס"ד
Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2

ISSN 1538 − 9472

Journal of **Modern Applied Statistical Methods**

JMASM (ISSN 1538−9472, http://digitalcommons.wayne.edu/jmasm) is an independent, open access electronic journal, published biannually in May and November by JMASM Inc. (PO Box 48023, Oak Park, MI, 48237) in collaboration with the Wayne State University Library System. *JMASM* seeks to publish (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo- random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal correspondence (other than manuscript submissions) and requests for advertising may be forwarded to ea@jmasm.com. See back matter for instructions for authors.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2

ISSN 1538 − 9472

Journal of Modern Applied Statistical Methods

Vol. 12, No. 2

NOVEMBER 2013

Table of Contents

Invited Articles

Regular Articles

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2

ISSN 1538 − 9472

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2

ISSN 1538 − 9472

Emerging Scholars

Statistical Software Applications and Review

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 2-19.

ISSN 1538 − 9472

Invited Article: Preliminary Testing for Normality: Is This a Good Practice?

H. J. Keselman
University of Manitoba
Winnipeg, Manitoba

Abdul R. Othman
Universiti Sains Malaysia
Georgetown, Penang

Rand R. Wilcox
University of S. California
Los Angeles, CA

Normality is a distributional requirement of classical test statistics. In order for the test statistic to provide valid results leading to sound and reliable conclusions this requirement must be satisfied. In the not too distant past, it was claimed that violations of normality would not likely jeopardize scientific findings (See Hsu & Feldt, 1969; Lunney, 1970). Recent revelations suggest otherwise (See e.g., Micceri, 1989; Keselman, Huberty, Lix et al., 1998; Erceg-Hurn, Wilcox, & Keselman, 2013; Wilcox and Keselman, 2003; Wilcox, 2012a, b). Unfortunately the data obtained in psychological investigations rarely, if ever, meet the requirement of normally distributed data (Micceri, 1989; Wilcox, 2012a, b). Consequently, it could be the case that the results from many of the investigations conducted in psychology provide invalid results. Accordingly, authors recommend that researchers attempt to assess the validity of assuming data are normal in form prior to conducting a test of significance (Erceg-Hurn, et al., 2013; Keselman, et al., 1998). Present evidence suggests that a popular fit-statistic, the Kolmogorov-Smirnov test does a poor job of evaluating whether data are normal. Our investigation based on this statistic and other fit-statistics provides a more favorable picture of preliminary testing for normality.

Keywords: Assessing normality, fit statistics, g-and-h non-normal skewed and kurtotic data, contaminated mixed-normal distributions; outlying value(s), Likert scales

Introduction

Psychological researchers are often reminded that the validity of their statistical tests and the conclusions derived from these tests depends to a great extent on whether the derivational assumptions of the test procedures have been satisfied (e.g., See Keselman, Huberty, Lix et al., 1998; Wilcox, 2012a, b; Wilcox &

H. J. Keselman is a Professor of Psychology. His research interests are in applied statistics. Email him at: kesel@ms.umanitoba.ca. Abdul R. Othman is a Professor of Statistics. Email him at: arothman60@yahoo.com. Rand R. Wilcox is a Professor of Psychology. Email him at: rwilcox@usc.edu.

Keselman, 2003). Consequently, though not a common practice, researchers are still reminded about assessing derivational assumptions (See Erceg-Hurn, Wilcox, & Keselman, 2013; Kirk, 2013; Schoder, Himmelmann & Wilhelm, 2006; Wilcox & Keselman, 2003). Almost all inferential methods require that in the population(s) the data is (are) normally distributed (as well as other requirements not relevant to this paper). Violation of the normality assumption can have a deleterious effect on the Type I error rate of test statistics (See Wilcox, 2012a, b; Wilcox & Keselman, 2003). Although the Type I error rate is widely viewed as being relatively unaffected by non-normality, Bradley (1980) has pointed out conditions in which this is not true. This finding is also evident in the findings of recent studies and published texts (e. g., See Hempel, Ronchetti, & Rousseeuw, 1986; Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006; Micceri, 1989; Schoder, et al., 2006; Staudte & Sheather, 1990; Wilcox, 2012a, b; Wilcox & Keselman, 2003).

Applied researchers can examine plots of their data and/or perform tests to assess the assumption, i. e., normality. Evaluating graphs (e.g., box-plots, stem-and-leaf, box and whisker, QQ plots) of ones data to assess whether data are normally distributed can be problematic since the determination relies on a subjective assessment (Wilk & Gnanadesikan, 1968). Thus, this practice is oftentimes not typically used when assessing the shape of the distribution of data (See Schoder, et al., 2006). Researchers tend to prefer exact methods based on formal tests for normality such as the Kolmogorov-Smirnov (K-S) goodness-of-fit statistic (See Muller & Fetterman, 2002, Chapter 7). Furthermore, researchers commonly use the result from a goodness-of-fit test to determine whether the normality of classical test procedures is satisfied thus providing legitimacy to the use of a classical test statistic. Consequently, preliminary testing for normality or any distributional shape is quite important in the whole inferential process and has been discussed in various contexts (See e.g., Cardoso de Oliveira & Ferreira, 2010; Doornik & Hansen, 2008; Sürücü, 2006). However, if the assumption of normality does not appear to be satisfied, researchers use this information to select alternative procedures such as nonparametric methods. Thus, it is important to know how well a preliminary test for normality, e. g., the K-S test, works in detecting non-normal data.

Unfortunately, according to Schoder, Himmelman, and Wilhelm (2006) "The Kolmogorov-Smirnov test performs badly on data with single outliers, 10% outliers, and skewed data at sample sizes <100." (p. 757) These authors investigated the performance of the K-S test for four types of non-normal data (e.g., normal distribution with a single outlier, normal distribution with 10%

outliers, skewed lognormal distribution with varying skewness, and an ordinal 5-point Likert scale with varying multinomial probabilities) and varying sample size in a pretest-posttest design. The assessment for normality was conducted at a 5% significance level. Unfortunately, the results tabled by Schoder et al. do not support the use of the K-S test as a preliminary test to assess normality of the data.

Because it is strongly believed that validity assumptions such as normality should be verified before adopting a classical test of significance that assumes the data in the population is normal in shape, it important to replicate the findings reported by Schoder, Himmelmann, and Wilhelm (2006) and extend their study in important ways. (For a contrary view previously noted in this journal, see Sawilowsky, 2002, p. 466-467). Other goodness-of-fit statistics are available (see, e.g., Muller & Fetterman, 2002, Chapter 7). Accordingly, a simulation study was conducted investigating three goodness-of-fit statistics, varying the degree of non-normality with other distributional shapes not investigated by Schoder, Himmelmann, and Wilhelm (2006), using sample sizes more likely to be encountered in psychological and educational research.

Method

Specifically, in this study the following are manipulated: (1) the procedure used to assess shape of distribution [K-S, Cramer-von Mises (CvM), Anderson-Darling (A-D)] fit-statistics (available through the SAS system), (2) the shapes of distributions (26 cases—14 g-and-h distributions, 8 contaminated normal mixture models, and 4 multinomial models), (3) the sample sizes (20, 40, and 80), depending on distribution, and (4) the level of significance for the fit-statistics (i.e., $\alpha = .05, .10, .15$ and $.20$) .

Most statistical packages (e.g., the SAS system) provide numerous fit statistics. Accordingly, it is possible that other tests other than the K-S can adequately assess whether normality, or other distributions, exists in the data. The SAS system was used to implement the K-S, CvM, and A-D fit-statistics. The choices of non-normal distributions are modifications from Schoder, Himmelmann and Wilhelm (2006) and Zimmerman (1998). Schoder, et al. (2006) investigated a normal distribution with a single outlier, a normal distribution with 10% outliers, skewed lognormal distributions with varying skewness, and an ordinal 5-point Likert scale with varying multinomial probabilities (common they state in dermatological investigations). Many non-normal distributions were investigated via g-and-h distributions (See Headrick, Kowalchuk, & Sheng, 2008; Hoaglin, 1983; 1985; Kowalchuk & Headrick, 2010; Tukey, 1960). These

distributions with their values for skewness and kurtosis are enumerated in Table 1. A substantial number of values of g and h were chosen to cover as broad a spectrum of non-normal distributions that could occur in psychological and behavioral science experiments (e.g., See Keselman, Huberty, Lix et al., 1998; Micceri, 1989; Wilcox, 2012a, b).

Table 1. g-and h-distributions examined in the simulation study with their corresponding measures of skewness and kurtosis

Distribution	Skewness	Kurtosis	Distribution	Skewness	Kurtosis
g=0,h=.05	0.00	0.82	g=.4,h=0	1.32	3.26
g=0,h=.075	0.00	1.49	g=.6,h=0	2.26	10.27
g=0,h=.1	0.00	2.51	g=1,h=0	6.19	110.94
g=0,h=.125	0.00	4.16	g=.2, h=.1	1.08	5.50
g=0,h=.15	0.00	7.17	g=.4,h=.1	2.45	20.30
g=0,h=.2	0.00	33.22	g=.6,h=.1	4.69	89.80
g=.2, h=0	0.61	0.68	g=.8,h=.1	9.27	603.61

Table 2. Contaminated mixed-normal distributions used in the power studies of the three goodness-of-fit-tests for normality

		Outliers	
n	Distribution	Distance (in standard Deviations)	Number
20	(.95)N(0,1) + (.05)N(0,25)	5	1
20	(.90)N(0,1) + (.10)N(0,25)	5	2
20	(.95)N(0,1) + (.05)N(0,100)	10	1
20	(.90)N(0,1) + (.10)N(0,100)	10	2
40	(.975)N(0,1) + (.025)N(0,25)	5	1
40	(.95)N(0,1) + (.05)N(0,25)	5	2
40	(.90)N(0,1) + (.10)N(0,25)	5	4
40	(.975)N(0,1) + (.025)N(0,100)	10	1
40	(.95)N(0,1) + (.05)N(0,100)	10	2
40	(.90)N(0,1) + (.10)N(0,100)	10	4
80	(.9875)N(0,1) + .(0125)N(0,25)	5	1
80	(.975)N(0,1) + (.025)N(0,25)	5	2
80	(.95)N(0,1) + (.05)N(0,25)	5	4
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1
80	(.975)N(0,1) + (.025)N(0,100)	10	2
80	(.95)N(0,1) + (.05)N(0,100)	10	4

The SAS system was used on a Sun Fire X4600 M2 x64 server: 8 x AMD Opteron Model 8220 processor (2.8GHz-dual-core) to generate g- and-h data, by modifying standard normal variates $Z \sim N(0,1)$ to non-normal variates by specifying values of g and h in the following quantile functions:

$$q(Z) = q_{g,h}(Z) = \frac{\exp(gZ) - 1}{g} \exp\left(\frac{hZ^2}{2}\right), \tag{1}$$

$$q(Z) = q_{g,0}(Z) = \frac{\exp(gZ) - 1}{g}, \tag{2}$$

$$q(Z) = q_{0,h(Z)} = Z \ \exp\left(\frac{hZ^2}{2}\right) \tag{3}$$

Equations (2) and (3) generate lognormal and symmetric h distributions, respectively. As Kowalchuk and Headrick (2010) noted "The parameter $\pm g$ controls the skew of a distribution in terms of both direction and magnitude. The parameter h controls the tail weight or elongation of a distribution and is positively related with kurtosis." (p. 63). As well, Type I error rates were investigated when data were obtained from a normal distribution [$g = h = 0$, the standard normal distribution (skewness and kurtosis = 0)].

A number of different contaminated mixed-normal distributions were examined, such as those reported in Zimmerman (1998). Contaminated mixed-normal distributions have one or more outlying values that deviate from the central mean of the distribution by some amount measured in standard deviation units. For example, Zimmerman examined a mixed normal distribution consisting of samples from $N(0,1)$ with probability .95 and from $N(0,400)$ with probability .05. Tukey (1960) suggested that outliers are a common occurrence in distributions and others have indicated that skewed distributions frequently depict psychological data (e.g., reaction time data). Accordingly, eight contaminated mixed normal distributions were examined that had one, two, or four outlying values which were five or ten standard deviations from the mean value. These distributions are enumerated in Table 2.

Finally, like Schoder, Himmelmann and Wilhelm (2006), a 5-point Likert scale was simulated; such data is frequently gathered in psychological (e.g., from clinical, personality, and social psychological investigations) and other behavioral science investigations. The same conditions investigated by Schoder et al. (2006) were investigated. Specifically,

1) even distribution (p=.02 for each category 0-4);
2) symmetric distribution
   ($p_0 = 0.1,\ p_1 = 0.2,\ p_2 = 0.4,\ p_3 = 0.2,\ p_4 = 0.1$) ;
3) moderately skewed distribution
   ($p_0 = 0.5,\ p_1 = 0.3,\ p_2 = 0.15,\ p_3 = 0.04,\ p_4 = 0.01$) ; and
4) heavily skewed distribution
   ($p_0 = 0.7,\ p_1 = 0.2,\ p_2 = 0.06,\ p_3 = 0.03,\ p_4 = 0.01$) .

Thus, for the 5-point Likert scale data there were 4 multinomial distributions that were simulated (See Table 3).

Table 3. Multinomial distributions based upon Schoder, Himmelmann, and Wilhelm's (2006) probabilities simulated as a five-point Likert Scale

	Even	Symmetric	Moderately Skewed	Heavily Skewed
n	(p_0,p_1,p_2,p_3,p_4)	(p_0,p_1,p_2,p_3,p_4)	(p_0,p_1,p_2,p_3,p_4)	(p_0,p_1,p_2,p_3,p_4)
20	(.2, .2, .2, .2, .2)	(.1, .2, .4. .2, .1)	(.5, .3, .15, .04, .01)	(.7, .2, .06, .03, .01)
40	(.2, .2, .2, .2, .2)	(.1, .2, .4. .2, .1)	(.5, .3, .15, .04, .01)	(.7, .2, .06, .03, .01)
80	(.2, .2, .2, .2, .2)	(.1, .2, .4. .2, .1)	(.5, .3, .15, .04, .01)	(.7, .2, .06, .03, .01)

The same number of sample size conditions as Schoder, Himmelmann, and Wilhelm (2006) were not investigated, but a reasonable range of values (i.e., n = 20,40,80) were includled, depending on the condition investigated. Specifically,

(i) for the 14 g- and h- distributions, and 2 contaminated normal distributions, .95N(0,1) + .05N(0, k), k=25, 100, sample sizes of 20, 40 and 80 were chosen.
(ii) For 2 contaminated normal distributions, .9N(0,1) + .1N(0,k), k=25, 100, sample sizes of 20 and 40 were chosen.
(iii) For 2 contaminated normal distributions, .975N(01) + .025N(0,k), k=25, 100, sample sizes of 40 and 80 were chosen.
(iv) For 2 contaminated normal distributions, .9875N(0,1) + .0125N(0,k), k=25, 100, sample size of 80 was chosen.

Lastly, because in preliminary testing it would be quite important to guard against a Type II error (falsely accepting the null hypothesis that the data are normal in form), we selected significance levels of .10 , .15, and .20, in addition to the standard .05. Each condition in the investigation was replicated 5,000 times.

Table 4. Power rates for the goodness-of-fit test on normality (n = 20).

	Distribution	Skewness	Kurtosis	α = .05	α = .10	α = .15
Kolmogorov-Smirnov	***Normal****	***0.00***	***0.00***	***0.0524***	***0.1082***	***0.1530***
	g=0,h=.05	0.00	0.82	0.0726	0.1304	0.1834
	g=0,h=.075	0.00	1.49	0.0870	0.1540	0.2094
	g=0,h=.1	0.00	2.51	0.1066	0.1838	0.2392
	g=0,h=.125	0.00	4.16	0.1320	0.2156	0.2726
	g=0,h=.15	0.00	7.17	0.1626	0.2502	0.3100
	g=0,h=.2	0.00	33.22	0.2296	0.3194	0.3834
	g=.2, h=0	0.61	0.68	0.1030	0.1678	0.2286
	g=.4,h=0	1.32	3.26	0.2436	0.3506	0.4262
	g=.6,h=0	2.26	10.27	0.4450	0.5662	0.6416
	g=1,h=0	6.19	110.94	0.7852	0.8648	0.9008
	g=.2, h=.1	1.08	5.50	0.1662	0.2530	0.3100
	g=.4,h=.1	2.45	20.30	0.3218	0.4204	0.4842
	g=.6,h=.1	4.69	89.80	0.5018	0.6026	0.6642
	g=.8,h=.1	9.27	603.61	0.6698	0.7602	0.8096
Cramer-von Mises	***Normal****	***0.00***	***0.00***	***0.0494***	***0.1036***	***0.1490***
	g=0,h=.05	0.00	0.82	0.0752	0.1368	0.1952
	g=0,h=.075	0.00	1.49	0.0970	0.1658	0.2260
	g=0,h=.1	0.00	2.51	0.1286	0.1996	0.2632
	g=0,h=.125	0.00	4.16	0.1608	0.2426	0.3038
	g=0,h=.15	0.00	7.17	0.1990	0.2842	0.3400
	g=0,h=.2	0.00	33.22	0.2756	0.3580	0.4232
	g=.2, h=0	0.61	0.68	0.1100	0.1814	0.2444
	g=.4,h=0	1.32	3.26	0.3082	0.4064	0.4872
	g=.6,h=0	2.26	10.27	0.5570	0.6590	0.7204
	g=1,h=0	6.19	110.94	0.8822	0.9268	0.9484
	g=.2, h=.1	1.08	5.50	0.1990	0.2826	0.3454
	g=.4,h=.1	2.45	20.30	0.3808	0.4730	0.5370
	g=.6,h=.1	4.69	89.80	0.5922	0.6728	0.7216
	g=.8,h=.1	9.27	603.61	0.7594	0.8552	0.8626
Anderson-Darling	***Normal****	***0.00***	***0.00***	***0.0494***	***0.1036***	***0.1490***
	g=0,h=.05	0.00	0.82	0.0810	0.1456	0.2040
	g=0,h=.075	0.00	1.49	0.1090	0.1816	0.2378
	g=0,h=.1	0.00	2.51	0.1444	0.2162	0.2766
	g=0,h=.125	0.00	4.16	0.1784	0.2582	0.3200
	g=0,h=.15	0.00	7.17	0.2182	0.2992	0.3590
	g=0,h=.2	0.00	33.22	0.2924	0.3798	0.4386
	g=.2, h=0	0.61	0.68	0.1222	0.1966	0.2584
	g=.4,h=0	1.32	3.26	0.3388	0.4456	0.5258
	g=.6,h=0	2.26	10.27	0.6012	0.6988	0.7528
	g=1,h=0	6.19	110.94	0.9086	0.9448	0.9602
	g=.2, h=.1	1.08	5.50	0.2190	0.2984	0.3610
	g=.4,h=.1	2.45	20.30	0.4084	0.4968	0.5590
	g=.6,h=.1	4.69	89.80	0.6168	0.6972	0.7444
	g=.8,h=.1	9.27	603.61	0.7876	0.8474	0.8766

*Type 1 error rates

Table 5. Power rates for the goodness-of-fit test on normality (n = 40).

	Distribution	Skewness	Kurtosis	α = .05	α = .10	α = .15
Kolmogorov-Smirnov	***Normal****	***0.00***	***0.00***	***0.0524***	***0.1082***	***0.1530***
	g=0,h=.05	0.00	0.82	0.0726	0.1304	0.1834
	g=0,h=.075	0.00	1.49	0.0870	0.1540	0.2094
	g=0,h=.1	0.00	2.51	0.1066	0.1838	0.2392
	g=0,h=.125	0.00	4.16	0.1320	0.2156	0.2726
	g=0,h=.15	0.00	7.17	0.1626	0.2502	0.3100
	g=0,h=.2	0.00	33.22	0.2296	0.3194	0.3834
	g=.2, h=0	0.61	0.68	0.1030	0.1678	0.2286
	g=.4,h=0	1.32	3.26	0.2436	0.3506	0.4262
	g=.6,h=0	2.26	10.27	0.4450	0.5662	0.6416
	g=1,h=0	6.19	110.94	0.7852	0.8648	0.9008
	g=.2, h=.1	1.08	5.50	0.1662	0.2530	0.3100
	g=.4,h=.1	2.45	20.30	0.3218	0.4204	0.4842
	g=.6,h=.1	4.69	89.80	0.5018	0.6026	0.6642
	g=.8,h=.1	9.27	603.61	0.6698	0.7602	0.8096
Cramer-von Mises	***Normal****	***0.00***	***0.00***	***0.0564***	***0.1068***	***0.1542***
	g=0,h=.05	0.00	0.82	0.0950	0.1622	0.2212
	g=0,h=.075	0.00	1.49	0.1332	0.2094	0.2746
	g=0,h=.1	0.00	2.51	0.1860	0.2692	0.3328
	g=0,h=.125	0.00	4.16	0.2448	0.3314	0.3996
	g=0,h=.15	0.00	7.17	0.3132	0.4012	0.4722
	g=0,h=.2	0.00	33.22	0.4490	0.5294	0.5908
	g=.2, h=0	0.61	0.68	0.1936	0.2786	0.3474
	g=.4,h=0	1.32	3.26	0.5528	0.6618	0.7352
	g=.6,h=0	2.26	10.27	0.8628	0.9116	0.9360
	g=1,h=0	6.19	110.94	0.9948	0.9980	0.9990
	g=.2, h=.1	1.08	5.50	0.3286	0.4220	0.4894
	g=.4,h=.1	2.45	20.30	0.6394	0.7218	0.7738
	g=.6,h=.1	4.69	89.80	0.8728	0.9120	0.9308
	g=.8,h=.1	9.27	603.61	0.9664	0.9798	0.9866
Anderson-Darling	***Normal****	***0.00***	***0.00***	***0.0564***	***0.1024***	***0.1556***
	g=0,h=.05	0.00	0.82	0.1036	0.1724	0.2326
	g=0,h=.075	0.00	1.49	0.1504	0.2278	0.2968
	g=0,h=.1	0.00	2.51	0.2082	0.2978	0.3612
	g=0,h=.125	0.00	4.16	0.2766	0.3678	0.4288
	g=0,h=.15	0.00	7.17	0.3460	0.4326	0.4960
	g=0,h=.2	0.00	33.22	0.4740	0.5620	0.6216
	g=.2, h=0	0.61	0.68	0.2130	0.3046	0.3750
	g=.4,h=0	1.32	3.26	0.6130	0.7160	0.7776
	g=.6,h=0	2.26	10.27	0.8946	0.9398	0.9586
	g=1,h=0	6.19	110.94	0.9974	0.9988	0.9998
	g=.2, h=.1	1.08	5.50	0.3556	0.4522	0.5194
	g=.4,h=.1	2.45	20.30	0.6698	0.7510	0.7956
	g=.6,h=.1	4.69	89.80	0.8958	0.9252	0.9444
	g=.8,h=.1	9.27	603.61	0.9738	0.9858	0.9890

*Type 1 error rates

Table 6. Power rates for the goodness-of-fit test on normality (n = 80).

	Distribution	Skewness	Kurtosis	α = .05	α = .10	α = .15
Kolmogorov-Smirnov	***Normal****	***0.00***	***0.00***	***0.0534***	***0.1082***	***0.1580***
	g=0,h=.025	0.00	0.35	0.0696	0.1314	0.1828
	g=0,h=.05	0.00	0.82	0.0968	0.1742	0.2252
	g=0,h=.075	0.00	1.49	0.1446	0.2318	0.2980
	g=0,h=.1	0.00	2.51	0.2114	0.3172	0.3928
	g=0,h=.125	0.00	4.16	0.3012	0.4194	0.4934
	g=0,h=.15	0.00	7.17	0.3950	0.5154	0.5958
	g=0,h=.2	0.00	33.22	0.5904	0.6938	0.7526
	g=0,h=.225	0.00	154.84	0.6736	0.7624	0.8098
	g=.2, h=0	0.61	0.68	0.2530	0.3758	0.4494
	g=.4,h=0	1.32	3.26	0.7334	0.8334	0.8792
	g=.6,h=0	2.26	10.27	0.9692	0.9872	0.9936
	g=1,h=0	6.19	110.94	1.0000	1.0000	1.0000
	g=.2, h=.1	1.08	5.50	0.4448	0.5604	0.6296
	g=.4,h=.1	2.45	20.30	0.8196	0.8870	0.9132
	g=.6,h=.1	4.69	89.80	0.9762	0.9882	0.9932
	g=.8,h=.1	9.27	603.61	0.9982	1.0000	1.0000
Cramer-von Mises	***Normal****	***0.00***	***0.00***	***0.0558***	***0.1030***	***0.1512***
	g=0,h=.05	0.00	0.82	0.1194	0.1872	0.2480
	g=0,h=.075	0.00	1.49	0.1896	0.2740	0.3442
	g=0,h=.1	0.00	2.51	0.2792	0.3834	0.4538
	g=0,h=.125	0.00	4.16	0.3912	0.4936	0.5570
	g=0,h=.15	0.00	7.17	0.5004	0.5990	0.6626
	g=0,h=.2	0.00	33.22	0.6914	0.7654	0.8128
	g=.2, h=0	0.61	0.68	0.3172	0.4314	0.5108
	g=.4,h=0	1.32	3.26	0.8526	0.9082	0.9372
	g=.6,h=0	2.26	10.27	0.9950	0.9980	0.9990
	g=1,h=0	6.19	110.94	1.0000	1.0000	1.0000
	g=.2, h=.1	1.08	5.50	0.5402	0.6346	0.6942
	g=.4,h=.1	2.45	20.30	0.8926	0.9302	0.9498
	g=.6,h=.1	4.69	89.80	0.9928	0.9968	0.9982
	g=.8,h=.1	9.27	603.61	1.0000	1.0000	1.0000
Anderson-Darling	***Normal****	***0.00***	***0.00***	***0.0548***	***0.1046***	***0.1526***
	g=0,h=.05	0.00	0.82	0.1316	0.2112	0.2694
	g=0,h=.075	0.00	1.49	0.2158	0.3046	0.3804
	g=0,h=.1	0.00	2.51	0.3196	0.4220	0.4946
	g=0,h=.125	0.00	4.16	0.4328	0.5290	0.5996
	g=0,h=.15	0.00	7.17	0.5420	0.6396	0.7004
	g=0,h=.2	0.00	33.22	0.7270	0.7960	0.8358
	g=.2, h=0	0.61	0.68	0.3606	0.4802	0.5604
	g=.4,h=0	1.32	3.26	0.8982	0.9430	0.9608
	g=.6,h=0	2.26	10.27	0.9976	0.9996	0.9998
	g=1,h=0	6.19	110.94	1.0000	1.0000	1.0000
	g=.2, h=.1	1.08	5.50	0.5816	0.6692	0.7234
	g=.4,h=.1	2.45	20.30	0.9104	0.9424	0.9608
	g=.6,h=.1	4.69	89.80	0.9942	0.9976	0.9986
	g=.8,h=.1	9.27	603.61	1.0000	1.0000	1.0000

*Type 1 error rates

Table 7. Number of times the g- and h- non-normal power values are equal to or greater than .80 for the fit-statistics (K-S, CvM, and A-D)

	n	K-S	CvM	A-D	
***α* = .05**	20	0	1	1	
	40	3	4	4	
	80	5	6	6	
	Total	**8**	**11**	**11**	**30**
***α* = .10**	20	1	2	2	
	40	4	4	4	
	80	6	7	8	
	Total	**11**	**13**	**14**	**38**
***α* = .15**	20	2	2	2	
	40	4	4	5	
	80	7	8	8	
	Total	**13**	**14**	**15**	**42**
***α* = .20**	20	**---**	2	2	
	40	**---**	5	6	
	80	**---**	8	8	
	Total	**---**	**15**	**16**	**31***
Grand Total			**53**	**56**	

Note: --- and *: PROC UNIVARIATE in SAS does not provide exact p-values for K-S at α = .20

Results

g- and h- Non-normal Distributions

Table 4 presents Type I error and power rates for the K-S, CvM, and A-D fit-statistics when sample size was 20. A number of conclusions can be drawn from this table. First, Type I error was controlled for each level of significance. Second, for the non-normal alternatives investigated, the K-S was typically the least powerful procedure, followed by CvM, and the A-D is typically most powerful. Also evident from the data is that for kurtotic data, none of the procedures displayed reasonable power (i.e., >.80). Although for skewed and kurtotic data the fit-statistics were only reasonably powerful for extreme departures from normality. As expected, power to detect non-normal distributions increased with more liberal

levels of significance; we excluded the $\alpha = .20$ values from the tables since the values are naturally larger than those reported for the other significance levels examined.

For moderate sample size case (See Table 5) the same pattern of results held; however, the fit-statistics had more power to detect non-normal data when sample size was 40. Finally, the same pattern of results occurred for our largest investigated sample size of 80 (See Table 6). And as expected the power to detect non-normal data increased with the increase in sample size.

To summarize the findings for the g-and-h non-normal distributions examined in this study we provide in Table 7 a count of the number of times the power values were equal to or greater than .80 across the simulated conditions. Over the significance levels that can be used with the K-S test (i.e., $\alpha = .05, .10,$ and $.15$) the A-D procedure was most powerful to detect non-normal distributions, followed closely by CvM and then by K-S. Clearly the A-D is most sensitive of the three. Also most evident is that the power to detect non-normal distributions is affected by the level of significance as would be expected. Also evident is that contrary to the warning given by Schoder, Himmelmann, and Wilhelm (2006) researchers can detect non-normal distributions with sample sizes less than 100 (80 in our case).

Contaminated Mixed-Normal Distributions

The power rates for the contaminated normal distributions for the three fit-statistics, K-S, CvM, and A-D are contained in Tables 8, 9, and 10, respectively. As we found for the g- and- h non-normal data, the A-D fit-statistic was most powerful for detecting normal data with outlying values than both the CvM and K-S fit-statistics. And, as expected, power increased with sample size and level of significance. Indeed, to a large extent the reported power values are in reasonably close proximity to .80 for most of the contaminated normal distributions examined. Furthermore, again, as expected the power values were largest when the level of significance was > .05.

Likert Non-normal data

The final type of non-normal data that we investigated was data that is obtained when five-point Likert scales are used in measuring the dependent variable. Subjects in the investigations indicate their preference, liking, attitude, etc. on five point type scales (e.g., very unfavorable, unfavorable, neutral, pleasant, very pleasant). Such responses obviously cannot be normally distributed.

Table 8. Power of the Kolmogorov-Smirnov goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α* =.05**	***α* =.10**	***α* =.15**
20	(.95)N(0,1) + (.05)N(0,25)	5	1	0.3238	0.4040	0.4568
20	(.9)N(0,1) + (.1)N(0,25)	5	2	0.4950	0.5748	0.6270
20	(.95)N(0,1) + (.05)N(0,100)	10	1	0.6022	0.6526	0.6866
20	(.9)N(0,1) + (.1)N(0,100)	10	2	0.8164	0.8566	0.8782
40	(.975)N(0,1) + (.025)N(0,25)	5	1	0.2898	0.3670	0.4240
40	(.95)N(0,1) + (.05)N(0,25)	5	2	0.4630	0.5424	0.5988
40	(.9)N(0,1) + (.1)N(0,25)	5	4	0.7050	0.7748	0.8144
40	(.975)N(0,1) + (.025)N(0,100)	10	1	0.5838	0.6462	0.6864
40	(.95)N(0,1) + (.05)N(0,100)	10	2	0.8160	0.8520	0.8732
40	(.9)N(0,1) + (.1)N(0,100)	10	4	0.9660	0.9768	0.9818
80	(.9875)N(0,1) + (.0125)N(0,25)	5	1	0.2472	0.3210	0.3804
80	(.975)N(0,1) + (.025)N(0,25)	5	2	0.4144	0.5006	0.5572
80	(.95)N(0,1) + (.05)N(0,25)	5	4	0.6754	0.7482	0.7852
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1	0.5436	0.6052	0.6464
80	(.975)N(0,1) + (.025)N(0,100)	10	2	0.7874	0.8288	0.8546
80	(.95)N(0,1) + (.05)N(0,100)	10	4	0.9606	0.9714	0.9778

Table 9. Power of the Cramer-von Mises goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α* =.05**	***α* =.10**	***α* =.15**
20	(.95)N(0,1) + (.05)N(0,25)	5	1	0.3692	0.4362	0.4844
20	(.9)N(0,1) + (.1)N(0,25)	5	2	0.5582	0.6220	0.6700
20	(.95)N(0,1) + (.05)N(0,100)	10	1	0.6386	0.6844	0.7164
20	(.9)N(0,1) + (.1)N(0,100)	10	2	0.8534	0.8802	0.8962
40	(.975)N(0,1) + (.025)N(0,25)	5	1	0.3374	0.4000	0.4590
40	(.95)N(0,1) + (.05)N(0,25)	5	2	0.5346	0.6018	0.6428
40	(.9)N(0,1) + (.1)N(0,25)	5	4	0.7776	0.8264	0.8540
40	(.975)N(0,1) + (.025)N(0,100)	10	1	0.6250	0.6698	0.7028

Table 9, continued. Power of the Cramer-von Mises goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α* =.05**	***α* =.10**	***α* =.15**
40	(.95)N(0,1) + (.05)N(0,100)	10	2	0.8478	0.8716	0.8864
40	(.9)N(0,1) + (.1)N(0,100)	10	4	0.9784	0.9836	0.9868
80	(.9875)N(0,1) + (.0125)N(0,25)	5	1	0.2928	0.3596	0.4150
80	(.975)N(0,1) + (.025)N(0,25)	5	2	0.4884	0.5548	0.6086
80	(.95)N(0,1) + (.05)N(0,25)	5	4	0.7534	0.8002	0.8298
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1	0.5924	0.6366	0.6714
80	(.975)N(0,1) + (.025)N(0,100)	10	2	0.8258	0.8580	0.8768
80	(.95)N(0,1) + (.05)N(0,100)	10	4	0.9736	0.9800	0.9836

Table 10. Power of the Anderson-Darling goodness-of-fit test on normality of data for contaminated normal distributions

		Outliers				
N	**Distribution**	**Distance (in std dev)**	**Number**	***α* =.05**	***α* =.10**	***α* =.15**
20	(.95)N(0,1) + (.05)N(0,25)	5	1	0.4024	0.4650	0.5136
20	(.9)N(0,1) + (.1)N(0,25)	5	2	0.5974	0.6596	0.6994
20	(.95)N(0,1) + (.05)N(0,100)	10	1	0.6688	0.7100	0.7368
20	(.9)N(0,1) + (.1)N(0,100)	10	2	0.8704	0.8922	0.9076
40	(.975)N(0,1) + (.025)N(0,25)	5	1	0.3802	0.4466	0.4944
40	(.95)N(0,1) + (.05)N(0,25)	5	2	0.5860	0.6432	0.6854
40	(.9)N(0,1) + (.1)N(0,25)	5	4	0.8174	0.8558	0.8762
40	(.975)N(0,1) + (.025)N(0,100)	10	1	0.6572	0.7000	0.7276
40	(.95)N(0,1) + (.05)N(0,100)	10	2	0.8664	0.8920	0.9056
40	(.9)N(0,1) + (.1)N(0,100)	10	4	0.9824	0.9866	0.9896
80	(.9875)N(0,1) + (.0125)N(0,25)	5	1	0.3356	0.4050	0.4576
80	(.975)N(0,1) + (.025)N(0,25)	5	2	0.5460	0.6138	0.6574
80	(.95)N(0,1) + (.05)N(0,25)	5	4	0.8036	0.8440	0.8694
80	(.9875)N(0,1) + (.0125)N(0,100)	10	1	0.6284	0.6726	0.7042
80	(.975)N(0,1) + (.025)N(0,100)	10	2	0.8568	0.8830	0.8990
80	(.95)N(0,1) + (.05)N(0,100)	10	4	0.9792	0.9850	0.9886

Table 11. Power of the goodness–of-fit test on normality of data for multinomial data representing five-point Likert scale scores

Kolmogoroc-Smirnov*

Even

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.2, .2, .2, .2, .2)	0.2742	0.4490	0.5630	
20	(.2, .2, .2, .2, .2)	0.6368	0.8248	0.8918	
40	(.2, .2, .2, .2, .2)	0.9978	1.0000	1.0000	

Symmetric

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.1, .2, .4, .2, .1)	0.4568	0.6164	0.7212	
20	(.1, .2, .4, .2, .1)	0.8132	0.9390	0.9606	
40	(.1, .2, .4, .2, .1)	0.9998	1.0000	1.0000	

Moderately Skewed

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.5, .3, .15, .04, .01)	0.7629	0.9176	0.9530	
20	(.5, .3, .15, .04, .01)	0.9970	0.9992	1.0000	
40	(.5, .3, .15, .04, .01)	1.0000	1.0000	1.0000	

Heavily Skewed

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.7, .2, .06, .03, .01)	0.9747	0.9905	0.9971	
20	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	
40	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	

Cramer-von Mises

Even

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.2, .2, .2, .2, .2)	0.2610	0.4000	0.5348	0.6620
20	(.2, .2, .2, .2, .2)	0.6710	0.9060	0.9946	1.0000
40	(.2, .2, .2, .2, .2)	0.9978	1.0000	1.0000	1.0000

Symmetric

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.1, .2, .4, .2, .1)	0.4520	0.5596	0.7008	0.7714
20	(.1, .2, .4, .2, .1)	0.8494	0.9664	0.9986	1.0000
40	(.1, .2, .4, .2, .1)	1.0000	1.0000	1.0000	1.0000

Moderately Skewed

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.5, .3, .15, .04, .01)	0.8501	0.9134	0.9734	0.9814
20	(.5, .3, .15, .04, .01)	0.9998	1.0000	1.0000	1.0000
40	(.5, .3, .15, .04, .01)	1.0000	1.0000	1.0000	1.0000

Heavily Skewed

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.7, .2, .06, .03, .01)	0.9825	0.9916	0.9981	0.9988
20	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	1.0000
40	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	1.0000

Anderson-Darling[b]

Even

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.2, .2, .2, .2, .2)	0.3202	0.5086	0.6220	0.7250
20	(.2, .2, .2, .2, .2)	0.8420	0.9888	1.0000	1.0000

Symmetric

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.1, .2, .4, .2, .1)	0.4248	0.5820	0.6628	0.7898
20	(.1, .2, .4, .2, .1)	0.8668	0.9916	1.0000	1.0000

Moderately Skewed

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.5, .3, .15, .04, .01)	0.8996	0.9526	0.9738	0.9928
20	(.5, .3, .15, .04, .01)	1.0000	1.0000	1.0000	1.0000

Heavily Skewed

n	(p_0,p_1,p_2,p_3,p_4)	α = .05	α = .10	α = .15	α = .20
10	(.7, .2, .06, .03, .01)	0.9897	0.9969	0.9988	0.9996
20	(.7, .2, .06, .03, .01)	1.0000	1.0000	1.0000	1.0000

*PROC UNIVARIATE does not allow α = .20 for the Kolmogorov-Smirnov test.
[b]The power values for non-tabled n = 40 values are all 1.000

Table 11 provides power rates for the three fit-statistics for detecting non-normality arising from using a Likert scale for assessing the dependent variable. Preliminary findings indicated that power values were 100% for sample sizes greater than 20 in the vast majority of cases. Thus, it was decided to include a smaller sample size case (i.e., $n = 10$) to examine power values for a relatively modest number of subjects. The findings are quite positive; that is, in just about every case examined, the power to detect non-normality is > .80. Indeed, out of the 106 tabled values 83 are greater in value than .80. Once again, the A-D statistic provides the best power values, followed by CvM, and then by K-S.

Discussion

Applied researchers use statistical tests to assess whether or not the effect of an experimental manipulation is significant. Unfortunately, the results of many of these investigations are suspect as they often involve the use of statistical procedures with questionable validity. In these cases, the reported effects may be misleading or, in many cases, wrong. Clearly, such erroneous decisions can have serious negative consequences for both the advancement of knowledge in a given field as well as the effective translation of research results into practice. The intent of this paper was to examine whether one can effectively test whether one's data confirms to the validity assumption of normality—a requirement for most classical test statistics. Prior research suggested that one could not use the Kolmogorov-Smirnov goodness-of-fit test to effectively test whether data were normally distributed or not (See e.g., Schoder, Himmelmann, and Wilhelm, 2006).

We looked into this negative finding by also investigating other fit statistics, the Cramer von Mises and Anderson-Darling tests (See Muller & Fetterman, 2002 Chapter 7), varying the skewness and kurtosis values of numerous g-and h-distributions, examining a number of contaminated mixed-normal distributions and examining results when the dependent variable was obtained from non-normal five-point Likert data. We also manipulated sample sizes (n = 20,40,80 and the level of significance for the test of normality $\alpha = .05, .10, .15$ and $.20$).

Of the three fit-statistics we found that the Anderson-Darling procedure was most effective in detecting non-normality being superior to both the Kolmogorov-Smirnov and Cramer-von Mises tests. We also determined that one could reasonably detect non-normality with reasonable sample sizes (n = 10,20,40), unlike what was reported by Schoder, Himmelmann, and Wilhelm (2006). Lastly, and importantly, since in this context one would want to increase the power to detect effects and concomitantly reduce the probability of falsely accepting the

null hypothesis that data are normally distributed, we suggest that preliminary testing be performed with significance levels larger than .05, say $\alpha = .15$ or $\alpha = .20$.

We conclude by reminding researchers that if normality is not present in the data current analytic practices allow researchers to test hypotheses say about mean equality in multiple group designs with software that does not require that data be normally distributed (See e. g., SAS's Glimmix procedure). Or, researchers can choose to replace classical test statistics and their least squares estimators for the mean and variance with robust test statistics with robust estimators (i.e., trimmed means and Winsorized variances (See e.g., Wilcox, 2012a, b; Wilcox & Keselman, 2003), procedures that have been found to be robust to non-normality [e.g., Erceg-Hurn, Wilcox, & Keselman (2013); Keselman, Algina, Lix, Wilcox, & Deering (2008a, b)].

References

Bradley, J. V. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society, 16*, 333-336.

Cardoso de Oliveira, I. R., & Ferreira, D. F. (2010). Multivariate extension of chi-squared univariate normality test. *Journal of Statistical Computation and Simulation, 80*, 513-526.

Doornik, J. A., & Hansen, H. (2008). Practioners' corner: An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70 supplement: 927-939.

Erceg-Hurn, D. M., Wilcox, R. R., & Keselman, H. J. (2013). Robust statistical estimation. In T. Little (Ed.), *The Oxford handbook of quantitative methods, Vol. 1*, 388-406. New York: Oxford University Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986).*Robust statistics*. New York: Wiley.

Headrick, T. C., Kowalchuk, R. K., & Sheng, Y. (2008). Parametric probability densities and distribution functions for Tukey g-and-h transformations and their use for fitting data. *Applied Mathematical Sciences, 2(9)*, 449-462.

Hoaglin, D. C. (1983). G-and-h distributions. In Kotz, S., & Johnson, N. L. Eds.), *Encyclopedia of statistical sciences*, Vol. 3, pp. 298-301. New York: Wiley.

Hoaglin, D. C. (1985). Summarizing shape numerically; The g-and h-distributions. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.), *Exploring data, tables, trends, and shapes*, pp. 461-511. New York: Wiley

Hsu, T., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. *American Educational Research Journal, 6*, 515-527.

Huber, P. J. & Ronchetti, E. (2009). *Robust statistics*, (2nd Ed.) New York: Wiley.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008a). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110-129.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008b). Supplemental materials to 129a. A SAS program to implement a general approximate degrees of freedom solution for inference and estimation. http://dx.doi.org/10.1037/1082-989X.13.2.110.supp

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Kirk, R. E. (2013). Experimental design: Procedures for the Behavioral Sciences (4^{th} ed). Los Angeles: Sage.

Kowalchuk, R. K., & Headrick, T. C. (2010). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology, 63*, 63-74.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement, 7*, 263-269.

Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006). *Robust statistics*: *Theory and methods*. New York: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Muller, K. E., & Fetterman, B. A. (2002). Regression and ANOVA: An integrated approach using SAS software. Cary, NC: SAS Institute, Inc.

SAS Institute. (2012). Statistics: ANOVA and regression. Cary, NC: SAS Institute, Inc.

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods*, *1*(2), 461-472.

Schoder, V., Himmelmann, A., & Wilhelm, K. P. (2006). Preliminary testing for normality: some statistical aspects of a common concept. *Clinical Dermatology*, *31*,757-761.

Staudte, R. G. & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Sürücü, B. (2006). Goodness-of-fit tests for multivariate distributions. *Communications in Statistics—Theory and Methods*, *35*, 1319-1331.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Wilcox, R. R. (2012a). *Introduction to robust estimation and hypothesis testing*, (3rd ed.) San Diego, CA: Academic Press.

Wilcox, R. R. (2012b). *Modern statistics for the social and behavioral sciences*: *A practical introduction*. New York: Chapman & Hall/CRC Press.

Wilcox, R. R. & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254-274.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*, 1-17.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education*, *67*, 55-68.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 20-34.

ISSN 1538 − 9472

Invited Article: Robust Regression Estimators When There are Tied Values

Rand R. Wilcox
University of Southern California
Los Angeles, CA

Florence Clark
University of Southern California
Los Angeles, CA

It is well known that when using the ordinary least squares regression estimator, outliers among the dependent variable can result in relatively poor power. Many robust regression estimators have been derived that address this problem, but the bulk of the results assume that the dependent variable is continuous. It is demonstrated that when there are tied values, several robust regression estimators can perform poorly in terms of controlling the Type I error probability, even with a large sample size. The presence of tied values does not necessarily mean that they perform poorly, but there is the issue of whether there is a robust estimator that performs reasonably well in situations where other estimators do not. The main result is that a modification of the Theil–Sen estimator achieves this goal. Results on the small-sample efficiency of the modified Theil–Sen estimator are reported as well. Data from the Well Elderly 2 Study, which motivated this study, are used to illustrate that the modified Theil–Sen estimator can make a practical difference.

Keywords: Tied values, Harrell–Davis estimator, MM–estimator, Coakley–Hettmansperger estimator, rank-based regression, Theil–Sen estimator, Well Elderly II Study, perceived control

Introduction

It is well known that the ordinary least squares (OLS) regression estimator is not robust (e.g., Hampel et al., 1987; Huber & Ronchetti, 2009; Maronna et al. 2006; Staudte & Sheather, 1990; Wilcox, 2012a, b). One concern is that even a single outlier among the values associated with the dependent variable can result in relatively poor power. Numerous robust regression estimators have been derived that are aimed at dealing with this issue, a fairly comprehensive list of which can be found in Wilcox (2012b, Chapter 10). But the bulk of the published

Rand R. Wilcox is a Professor of Psychology. Email him at: rwilcox@usc.edu. Florence Clark is a Professor of Occupational Science and Occupational Therapy. Email her at: fclark@osot.usc.edu.

results on robust regression estimators assume the dependent variable is continuous.

Motivated by data stemming from the Well II study (Jackson et al. 2009), this paper examines the impact of tied values on the probability of a Type I error when testing hypotheses via various robust regression estimators. Many of the dependent variables in the Well Elderly study were the sum of Likert scales. Consequently, with a sample size of 460, tied values were inevitable. Moreover, the dependent variables were found to have outliers, suggesting that power might be better using a robust estimator. But given the goal of testing the hypothesis of a zero slope, it was unclear whether the presence of tied values might impact power and the probability of a Type I error.

Preliminary simulations indicated that indeed there is a practical concern. Consider, for example, the Theil (1950) and Sen (1968) estimator. One of the dependent variables (CESD) in the Well Elderly study reflected a measure of depressive symptoms. It consists of the sum of twenty Likert scales with possible scores ranging between 0 and 60. The actual range of scores in the study was 0 to 56. Using the so-called MAD-median rule (e.g., Wilcox, 2012b), 5.9% of the values were flagged as outliers, raising concerns about power despite the relatively large sample size. A simulation was run where observations were randomly sampled with replacement from the CESD scores and the independent variable was taken to be values randomly sampled from a standard normal distribution and independent of the CESD scores. The estimated Type I error probability, when testing at the .05 level, was .002 based on 2000 replications. A similar result was obtained when the dependent variable was a measure of perceived control. Now 7.8% of the values are declared outliers. As an additional check, the values for the dependent variable were generated from a beta-binomial distribution having probability function

$$P(Y=y)=\frac{B(m-y+r,y+s)}{(m+1)B(m-y+1,y+1)B(r,s)}, \quad (1)$$

where B is the complete beta function and the sample space consists of the integers $0,\ldots,m$. For $r = s = 1$ as well as $(r, s) = (1, 9)$, again, the actual level was less than .01.

Other robust estimators were found to have a similar problem or situations were encountered where they could not be computed. The estimators that were considered included Yohai's (1987) MM-estimator, the one-step estimator derived

by Agostinelli and Markatou (1998), Rousseeuw's (1984) least trimmed squares (LTS) estimator, the Coakley and Hettmansperger (1993) M-estimator, the Koenker and Bassett (1978) quantile estimator and a rank-based estimator stemming from Jaeckel (1972). The MM-estimator and the LTS estimator were applied via the R package robustbase, the Agostinelli—Markatou estimator was applied with the R package wle, the quantile regression estimator was applied via the R package quantreg, the rank-based estimator was applied using the R package Rfit, and the Coakley–Hettmansperger and Theil—Sen estimators were applied via the R package WRS. A percentile bootstrap method was used to test the hypothesis of a zero slope, which allows heteroscedasticity and has been found to perform relatively well, in terms of controlling the probability of a Type I error, compared to other strategies that have been studied (Wilcox, 2012b). The MM-estimator, the Agostinelli—Markatou estimator and the Coakley—Hettmansperger estimator routinely terminated in certain situations due to some computational issue. This is not to suggest that they always performed poorly, this is not the case. But when dealing a skewed discrete distribution (a beta-binomial distribution with $m = 10$, $r = 9$ and $s = 1$), typically a p-value could not be computed. The other estimators had estimated Type I errors well below the nominal level. The R package Rfit includes a non-bootstrap test of the hypothesis that the slope is zero. Again the actual level was found to be substantially less than the nominal level in various situations, and increasing n only made matters worse. So this raised the issue of whether any reasonably robust estimator can be found that avoids the problems just described.

For completeness, when dealing with discrete distributions, an alternative approach is to use multinomial logistic regression. This addresses an issue that is potentially interesting and useful. But in the Well study, for example, what was deemed more relevant was modeling the typical CESD score given a value for CAR. That is, a regression estimator that focuses on some conditional measure of location, given a value for the independent variable, was needed.

The goal in this paper is to suggest a simple modification of the Theil–Sen estimator that avoids the problems just indicated. Section 2 reviews the Theil–Sen estimator and indicates why it can be highly unsatisfactory. Then the proposed modification is described. Section 3 describes the hypothesis testing method that is used. Section 4 summarizes simulation estimates of the actual Type I error probability when testing at the .05 level and it reports some results on its small-sample efficiency. Section 5 uses data from Well Elderly II study to illustrate that the modified Theil–Sen estimator can make a substantial practical difference.

The Theil–Sen Estimator and the Suggested Modification

When the dependent variable is continuous, the Theil–Sen estimator enjoys good theoretical properties and it performs well in simulations in terms of power and Type I error probabilities when testing hypotheses about the slope (e.g., Wilcox, 2012b). Its mean squared error and small-sample efficiency compare well to the OLS estimator as well as other robust estimators that have been derived (Dietz, 1987; Wilcox, 1998). Dietz (1989) established that its asymptotic breakdown point is approximately .29. Roughly, about 29% of the points must be changed in order to make the estimate of the slope arbitrarily large or small. Other asymptotic properties have been studied by Wang (2005) and Peng et al. (2008). Akritas et al. (1995) applied it to astronomical data and Fernandes and Leblanc (2005) to remote sensing. Although the bulk of the results on the Theil–Sen estimator deal with situations where the dependent variable is continuous, an exception is the paper by Peng et al. (2008) that includes results when dealing a discontinuous error term. They show that when the distribution of the error term is discontinuous, the Theil–Sen estimator can be super-efficient. They establish that even in the continuous case, the slope estimator may or may not be asymptotically normal. Peng et al. also establish the strong consistency and the asymptotic distribution of the Theil–Sen estimator for a general error distribution. Currently, a basic percentile bootstrap seems best when testing hypotheses about the slope and intercept, which has been found to perform well even when the error term is heteroscedastic (e.g., Wilcox, 2012b).

The Theil–Sen estimate of the slope is the usual sample median based on all of the slopes associated with any two distinct points. Consequently, practical concerns previously outlined are not surprising in light of results when dealing with inferential methods based on the sample median (Wilcox, 2012a, section 4.10.4). Roughly, when there are tied values, the sample median is not asymptotically normal. Rather, as sample size increases, the cardinality of its sample can decrease, which in turn creates concerns about the more obvious methods for testing hypotheses

Recent results on comparing quantiles (Wilcox et al., 2013) suggest a modification that might deal the concerns previously indicated: replace the usual sample median with the Harrell and Davis (1982) estimate of the median, which uses a weighted average of all the order statistics.

To describe the computational details, let (Y_1, X_1), …, (Y_n, X_n) be a random sample from some unknown bivariate distribution. Assuming that $X_j \neq X_k$ for any $j < k$, let

$$b_{jk} = \frac{Y_j - Y_k}{X_j - X_k}, 1 \le j < k \le n.$$

The Theil–Sen estimate of the slope, $\hat{\beta}_1$, is taken to be the usual sample median based on the b_{jk} values. The intercept is typically estimated with $\hat{\beta}_0 = M_y - \hat{\beta}_1 M_x$, where M_y is the usual sample median based on $Y_1, \ldots, Y_n$. This will be called the TS estimator henceforth.

For notational convenience, Let $Z_1, \ldots, Z_\ell$ denote the b_{jk} values, where $\ell = (n^2 - n)/2$. Let U be a random variable having a beta distribution with parameters $a = (\ell + 1)q$ and $b = (\ell + 1)(1 - q)$, $0 < q < 1$. Let

$$W_I = P\left(\frac{i-1}{\ell} \le U \le \frac{i}{\ell}\right).$$

Let $Z_{(1)} \le \ldots \le Z_{(\ell)}$ denote the $Z_1, \ldots, Z_\ell$ values written in ascending order. The Harrell and Davis (1982) estimate of the qth quantile is

$$\hat{\theta}_q = \sum W_i Z_{(i)}$$

Consequently, estimate the slope with $\tilde{\beta}_1 = \hat{\theta}_{.5}$. The intercept is estimated with the Harrell–Davis estimate of the median based on $Y_1 - \tilde{\beta}_1 X_1, \ldots, Y_n - \tilde{\beta}_n X_n$. This will be called the HD estimator.

So the strategy is to avoid the problem associated with the usual sample median by using a quantile estimator that results in a sampling distribution that in general does not have tied values. Because the Harrell–Davis estimator uses all of the order statistics, the expectation is that in general it accomplishes this goal. For the situations described in the introduction, for example, no tied values were found among the 5000 estimates of the slope. This, in turn, offers some hope that good control over the probability of a Type I error can be achieved via a percentile bootstrap method.

It is noted that alternative quantile estimators have been proposed that are also based on a weighted average of all the order statistics. In terms of its standard error, Sfakianakis and Verginis (2006) show that in some situations the Harrell–Davis estimator competes well with alternative estimators that again use a

weighted average of all the order statistics, but there are exceptions. Additional comparisons of various estimators are reported by Parrish (1990), Sheather and Marron (1990), as well as Dielman, Lowry and Pfaffenberger (1994). Perhaps one of these alternative estimators offers some practical advantage for the situation at hand, but this is not pursued here.

Hypothesis Testing

As previously indicated, a percentile bootstrap method has been found to be an effective way of testing hypotheses based on a robust regression estimators, including situations where the error term is heteroscedastic (e.g., Wilcox, 2012b). Also, because it is unclear when the HD estimator is asymptotically normal, using a percentile bootstrap method for the situation at hand seems preferable compared to using some pivotal test statistic based on some estimate of the standard error. (For general theoretical results on the percentile bootstrap method that are relevant here, see Liu & Singh, 1997.)

When testing

$$H_0 : \beta_1 = 0, \tag{2}$$

the percentile bootstrap begins by resampling with replacement n vectors of observations from $(Y_1, X_1), \ldots, (Y_n, X_n)$ yielding say $(Y_1^*, X_1^*), \ldots, (Y_n^*, X_n^*)$. Based on this bootstrap sample, let $\tilde{\beta}_1^*$ be the resulting estimate of the slope. Repeat this process B times yielding $\tilde{\beta}_{1b}^*, 1, \ldots, B$. Let A be the proportion of $\tilde{\beta}_{1b}^*$ values that are less than null value, 0, and let C be the number of times $\tilde{\beta}_{1b}^*$ is equal to the null value. Then a (generalized) p-value when testing (2) is

$$p = 2\min(\hat{p}, 1 - \hat{p}),$$

where $\hat{p} = \frac{A}{B} + .5\frac{C}{B}$. Here, B = 599 is used. This choice appears to work well with robust estimators in terms of controlling the probability of a Type I error (e.g., Wilcox, 2012b). However, based on results in Racine and MacKinnon (2007), $B > 599$ might provide improved power.

Simulation Results

Simulations were used to study the small-sample properties of the HD estimator. When comparing the small-sample efficiency of estimators, 4000 replications were used with $n = 20$. When estimating the actual probability of a Type I error, 2000 replications were used with sample sizes 20 and 60. Some additional simulations were run with $n = 200$ as a partial check on the R functions that were used to apply the methods.

To ensure tied values, values for Y were generated from one of four discrete distributions. The first two were beta-binomial distributions. Here $m = 10$ is used in which case the possible values for Y are the integers 0, 1, …, 10. The idea is to consider a situation where the number of tied values is relatively large. The values for r and s were taken to be $(r,s) = (1,9)$, which is a skewed distribution with mean 1, and $r = s = 3$, which is a symmetric distribution with mean 5. The third distribution was a discretized version of the normal distribution. More precisely, n observations were generated from a standard normal distribution, say $V_1,\ldots V_n$, and Y_i is taken to be $2V_i$ rounded to the nearest integer. (Among the 4,000 replications, the observed values for Y ranged between -9 and 10.) This process for generating observations will be labeled SN. For the final distribution, observations were generated as done in SN but with a standard normal replace by a contaminated normal having distribution

$$H(y) = .9\Phi(y) + .1\Phi\left(\frac{y}{10}\right),$$

where $\Phi(y)$ is a is a standard normal distribution. The contaminated normal has mean zero and variance 10.9. It is heavy-tailed, roughly meaning that it tends to generate more outliers than the normal distribution. This process will be labeled CN.

Estimated Type I error probabilities are shown in Table 1 for $n = 20$ and 60 when testing at the $\alpha = .05$ level. In Table 1, $B(r,s,m)$ indicates that Y has a beta-binomial distribution. The column headed by TS shows the results when using the Theil–Sen estimator. Notice that the estimates are substantially less than the nominal level when $n = 20$. Moreover, the estimated level actually decreases when n is increased to 60. In contrast, when using the HD estimator, the estimated level is fairly close to the nominal level among all of the situations considered, the estimates ranging between .044 and .057.

Negative implications about power seem evident when using TS. As a brief illustration, suppose that data are generated from the model $V = .25X + \varepsilon$, where X and ε are independent and both have a standard normal distribution. Let $Y = 2V$, rounded to the nearest integer. With $n = 60$, power based on TS was estimated to be .073. Using instead HD, power was estimated to be .40.

Table 1. Estimated Type I error probabilities, $\alpha = .05$

Distribution	n	TS	HD
B(3,3,10)	20	0.019	0.044
B(3,3,10)	60	0.002	0.047
B(1,9,10)	20	0.000	0.045
B(1,9,10)	60	0.000	0.045
SN	20	0.011	0.044
SN	60	0.001	0.050
CN	20	0.012	0.057
CN	60	0.004	0.048

Table 2. Estimated Efficiency, $n = 20$

Distribution	TS	TD
SN	0.809	1.090
B(3,3,10)	0.733	0.997
B(1,9,10)	0.689	2.610
CN	2.423	2.487

Of course, when Y has a discrete the least squares estimator could be used. To gain some insight into the relative merits of the HD estimator, its small-sample efficiency was compared to the least squares estimator and the TS estimator for the same situations in Table 1. Let V_0^2 be the estimated squared standard error of least squares estimate of the slope based on 4000 replications. Let V_1^2 and V_2^2 be the estimated squared standard errors for TS and HD, respectively. Then the efficiency associated with TS and HD was estimated with V_0 / V_1 and V_0 / V_2, respectively, the ratio of the estimated standard errors. Table 2 summarizes the results. As can be seen, the HD estimator competes very well with the least squares estimator. Moreover, there is no indication that TS ever

offers much of an advantage over HD, but HD does offer a distinct advantage over TS in some situations.

A related issue is the efficiency of the HD estimator when dealing with a continuous error term, including situations where there is heteroscedasticity. To address this issue, additional simulations were run by generating data from the model $Y = \lambda(X)\varepsilon$ where ε is some random variable having median zero and the function $\lambda(X)$ is used to model heteroscedasticity. The error term was taken to have one of four distributions: normal, symmetric with heavy tails, asymmetric with light tails and asymmetric with heavy tails. More precisely, the error term was taken to have a g-and-h distribution (Hoaglin, 1985) that contains the standard normal distribution as a special case. If Z has a standard normal distribution, then

$$W = \frac{\exp(gZ)\text{-}1}{g}\exp\left(hZ^2\right), \text{if } g > 0$$

and

$$W = Z\exp\left(h\frac{Z^2}{2}\right), \text{if } g = 0$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. As is evident, $g = h = 0$ corresponds to a standard normal distribution. Table 3 indicates the skewness (κ_1) and kurtosis (κ_2) of the four distributions that were used.

Table 3. Some properties of the g-and-h distribution

g	h	κ_1	κ_2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

Three choices for λ were used: $\lambda(X)=1$ (homoscedasticity), $\lambda(X)=|X|+1$ and $\lambda(X)=1/(|X|+1)$. For convenience, these three choices are denoted by variance patterns (VP) 1, 2, and 3.

Table 4 reports the estimated efficiency of TS and HD when *X* has a normal distribution. To provide a broader perspective, included are the estimated efficiencies of Yohai's (1987) MM-estimator and the least trimmed squares (LTS) estimator. Yohai's estimator was chosen because it has excellent theoretical properties. It has the highest possible breakdown point, .5, and it plays a central role in the robust methods discussed by Heritier et al. (2009). Both the MM-estimator and the LTS estimator were applied via the R package robustbase. As can be seen, for the continuous case, there is little separating the TS, HD and MM estimators with TS and MM providing a slight advantage over HD.

Table 4. Estimated efficiencies, the continuous case, *X* normal

g	h	VP	TS	HD	MM	LTS
0.000	0.000	1.000	0.861	0.930	0.967	0.708
		2.000	0.994	0.991	1.019	0.769
		0.300	0.997	0.966	0.999	0.776
0.000	0.200	1.000	1.234	1.157	1.199	0.971
		2.000	1.405	1.230	1.267	1.070
		3.000	1.389	1.216	1.276	1.041
0.200	0.000	1.000	0.897	1.146	0.960	0.989
		2.000	1.019	1.009	1.051	0.815
		3.000	0.978	0.999	1.026	0.793
0.200	0.200	1.000	1.314	1.200	1.259	1.022
		2.000	1.615	1.440	1.475	1.197
		3.000	1.443	1.271	1.337	1.160

There are situations where the differences in efficiency are more striking than those reported in Table 4. Also, no single estimator dominates in terms of efficiency: situations can be constructed where each estimator performs better than the others considered here. Suppose, for example, that *X* has a contaminated normal distribution and *Y* has a normal distribution. From basic principles, this situation favors OLS because as the distribution of *X* moves toward a heavy-tailed distribution, the standard error of the OLS estimator decreases. The resulting

efficiencies were estimated to be 0.514, 0.798, 0.844 and 0.533 for TS, HD, MM and LTS, respectively, with TS and LTS being the least satisfactory. Removing leverage points (outliers among the independent variable) using the MAD-median rule (e.g., Wilcox, 2012a, section 3.13.4), the estimates are 1.336, 1.727, 1.613 and 2.1213. So now LTS performs best in contrast to all of the other situations previously reported.

There is the issue of whether the MM-estimator has good efficiency for the discrete case. For the beta-binomial distribution with $r = s = 3$, the efficiency of the HD estimator is a bit better, but for the other discrete distributions considered here, the efficiency of the MM-estimator could not be estimated because the R function used to compute the MM-estimator routinely terminated with an error. For the same reason, the Type I error probability based on the hypothesis testing method used by the R package robustbase could not be studied. Switching to the bootstrap method used here only makes matters worse: bootstrap samples result in situations where the MM-estimator cannot be computed.

An Illustration

Using data from the Well Elderly II study (Jackson et al., 2009), it is illustrated that the choice between the TS and HD estimators can make a practical difference. A general goal in the Well Elderly II study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was aimed at understanding the association between the cortisol awakening response (CAR), which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep, and a measure of Perceived Control (PC), which is the sum of 8 four-point Likert scales. So the possible PC scores range between 8 and 32. Higher PC scores reflect greater perceived control. (For a detailed study of this measure of perceived control, see Eizenman et al., 1997.) CAR is taken to be the cortisol level upon awakening minus the level of cortisol 30-60 minutes after awakening.) Approximately 8% of the PC scores are flagged as outliers using the MAD-median rule. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicate that various forms of stress are associated with the CAR. After intervention, the TS estimate of the slope is -0.72 with a p-value of .34. Using instead HD, the estimate of the slope is -0.73 with a p-value less than .001.

Concluding Remarks

In summary, when dealing with tied values among the dependent variable, several robust estimators can result in poor control over the Type I error probability and relatively low power, so they should be used with caution. Moreover, the performance of the Theil–Sen estimator can actually deteriorate as the sample size increases. One way of dealing with this problem is to use the HD estimator, which is simple modification of the Theil–Sen estimator. In some situations the HD estimator has better efficiency than other robust estimators, but situations are encountered where the reverse is true. The very presence of tied values does not necessarily mean that robust estimators other than HD will perform poorly. The only point is that when dealing with tied values, the HD estimator can be computed in situations where other robust estimators cannot and it can provide a practical advantage in terms of both Type I error probabilities and power.

Various suggestions have been made about how to extend the Theil–Sen estimator to more than one independent variable (Wilcox, 2012b). One approach is the back-fitting algorithm, which is readily used in conjunction with the HD estimator. Here, the details are not of direct relevance so for brevity they are not provided. An R function, tshdreg, has been added to the R package WRS that performs the calculations.

References

Akritas, M. G., Murphy, S. A. & LaValley, M. P. (1995). The Theil–Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association 90,* 170-177.

Agostinelli, C. & Markatou, M. (1998) A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. *Statistics Probability Letters, 37,* 341-350.

Chida, Y. & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology, 80,* 265-278 .

Clow, A., Thorn, L., Evans, P. & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress, 7,* 29-37.

Coakley, C. W. & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association, 88,* 872-880.

Dielman, T., Lowry, C. & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics–Simulation and Computation, 23*, 355-371.

Dietz, E. J. (1987). A comparison of robust estimators in simple linear regression. *Communications in Statistics–Simulation and Computation, 16*, 1209-1227.

Dietz, E. J. (1989). Teaching regression in a nonparametric statistics course. *American Statistician, 43*, 35-40.

Eizenman, D. R., Nesselroade, J. R., Featherman, D. L. & Rowe, J. W. (1997). Intraindividual variability in perceived control in an older sample: The MacArthur successful aging studies. *Psychology and Aging, 12*, 489–502.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). Robust Statistics. New York: Wiley.

Fernandes, R. & Leblanc, S. G. (2005). Parametric (modified least squares) and non-parametric (Theil-Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment, 95*, 303-316.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69*, 635-640.

Heritier, S., Cantoni, E, Copt, S. & Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. New York: Wiley.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring Data Tables Trends and Shape*s. New York: Wiley, pp. 461-515.

Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics*, 2nd Ed. New York: Wiley.

Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., Chou, C.-P., Jordan-Marsh, M., Forman, T., White, B., Granger, D., Knight, B., & Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials, 6*, 90-101.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Annals of Mathematical Statistics, 43*, 1449-1458.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrika, 46*, 33-50.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association, 92*, 266-277.

Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006). Robust Statistics: Theory and Methods. New York: Wiley.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics, 46*, 247-257.

Peng, H., Wang, S. & Wang, X. (2008).Consistency and asymptotic distribution of the Theil- Sen estimator. Journal of Statistical Planning and Inference, 138, 1836-1850.

Racine, J. & MacKinnon, J. G. (2007). Simulation-based tests than can use any number of simulations. *Communications in Statistics–Simulation and Computation, 36*, 357-365.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*, 871-880.

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*, 1379-1389.

Sfakianakis, M. E. & Verginis, D. G. (2006). A new family of nonparametric quantile estimators. *Communications in Statistics–Simulation and Computation, 37*, 337-345.

Sheather, S. J. & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association, 85*, 410-416.

Staudte, R. G. & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae, 12*, 85-91.

Wang, X. Q. (2005). Asymptotics of the Theil-Sen estimator in simple linear regression models with a random covariate. *Nonparametric Statistics 17*, 107-120.

Wilcox, R. R. (1998). Simulation results on extensions of the Theil-Sen regression estimator. *Communications in Statistics–Simulation and Computation, 27*, 1117-1126.

Wilcox, R. R. (2012a). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. New York: Chapman Hall/CRC press.

Wilcox, R. R. (2012b). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd Edition. San Diego, CA: Academic Press.

Wilcox, R. R., Erceg-Hurn, D., Clark, F. Carlson, M. (2013). Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*. DOI: 10.1080/00949655.2012.754026

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics, 15*, 642-656.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 35-81.

ISSN 1538 − 9472

Regular Articles: A Monte Carlo Comparison of Robust MANOVA Test Statistics

Holmes Finch
Ball State University
Muncie, IN

Brian French
Washington State University
Pullman, WA

Multivariate Analysis of Variance (MANOVA) is a popular statistical tool in the social sciences, allowing for the comparison of mean vectors across groups. MANOVA rests on three primary assumptions regarding the population: (a) multivariate normality, (b) equality of group population covariance matrices and (c) independence of errors. When these assumptions are violated, MANOVA does not perform well with respect to Type I error and power. There are several alternative test statistics that can be considered including robust statistics and the use of the structural equation modeling (SEM) framework. This simulation study focused on comparing the performance of the *P* test statistics with fifteen other test statistics across seven manipulated factors. These statistics were evaluated across 12,076 different conditions in terms of Type I error and power. Results suggest that when assumptions were met, the standard MANOVA test functioned well. However, when assumptions were violated, it performed poorly, whereas several of the alternatives performed better. Discussion focuses on advice for selecting alternatives in practice. This study's focus on all these in one simulation and the 3 group case should be helpful to the practitioner making methodological sections.

Keywords: MANOVA, robust statistics, structural equation modeling, nonparametric, mean comparisons, Monte Carlo simulation

Introduction

Much research in the social sciences involves the comparison of means for two or more groups across multiple related outcome measures. For example, studies examining the impact of interventions on multiple measures of academic, social, communication, and emotional development are common in education and psychology. Parenting our Children to Excellence (*PACE*) (Dumas et al., 1999) is

Dr. Finch is a Professor of psychometrics and statistics in the Department of Educational Psychology. Email him at: whfinch@bsu.edu. Dr. French is a professor of measurement and psychometrics in the Department of Educational Leadership and Counseling Psychology. Email him at: frenchb@wsu.edu.

such an intervention project that has been tested through randomized control trials evaluating an 8-week program that teaches positive parenting techniques aimed at increasing parenting skills and child positive behavior. In programs such as this, there are typically multiple correlated outcome variables (e.g., child disruptive behaviors, child adjustment, parenting behaviors, parenting competence), which can have high-stakes implications (e.g., resource allocation, curriculum development, policy decisions). Therefore, given that high stakes decisions may be based upon the results of statistical analyses, precise modeling of data is paramount.

This type of research design in intervention work may revolve around hypotheses regarding group differences on a set of variables, rather than on individual variables. Multivariate hypotheses lead a researcher to a multivariate analysis, as it may be most appropriate for assessing group differences on the set of variables (Huberty & Olejnik, 2006). Specifically, multivariate analysis of variance (MANOVA) is well-suited for testing hypotheses about differences between groups (Hair, Anderson, Tatham, & Black, 1987). MANOVA can be viewed as a direct extension of the univariate general linear model that is most appropriate for examining differences between groups on several variables simultaneously (Hair et al., 1987; Olejnik, 2010). As Hancock, Lawrence and Nevitt (2001) pointed out, “MANOVA evaluates group differences on a linear composite of observed variables constructed so as to maximally differentiate the groups in multivariate space" (p. 535).

Situations are described here in which MANOVA may be the optimal analysis (particularly when compared with univariate analysis of variance (ANOVA)). Following this discussion, particular data structures that may cause problems for MANOVA will be described, particularly when key assumptions are violated, and then several approaches for dealing with the assumption violations. A simulation study comparing these methods across a variety of conditions is reported, and conclude the discussion with recommendations for researchers using MANOVA in cases where the assumptions are not met.

Despite the fact that MANOVA may be the optimal analysis for a multivariate problem due to its relative ease of use and interpretation, researchers may often employ multiple independent ANOVA models to determine if there are significant differences among group means on each of several outcome measures of interest. In the previous example with PACE, five separate ANOVAs could be conducted to determine if the treatment and control groups differed on the related outcomes. Although this approach may be familiar to many researchers, the simplicity of the univariate ANOVA could also lead to unwarranted conclusions

due to inflation of the family-wise Type I error rate and a potential decrease in power when the response is actually multivariate in nature. In fact, McCarroll, Crays, & Dunlap (1992) provided evidence that Type I error rates are inflated when ANOVA is used in a sequential manner. For example, the family-wise Type I error rate for testing the 5 outcomes in the PACE data, assuming alpha = 0.05, would be 0.23. It is acknowledged that by adjusting critical values for the univariate situation, the Type I error rate can be controlled (Ramsey, 1982). In fact, Ramsey illustrated that the Bonferroni procedures showed greater robustness in many cases compared to methods based on Hotelling's T^2 statistic, which requires more and stronger assumptions (e.g., multivariate normality) compared to Bonferroni procedures.

Often the research question of interest concerns differences on a set of related or correlated outcome variables, not each variable separately. That is, the researcher wants to examine questions about how groups differ along a combination of correlated dimensions or variables, not one dimension or variable at a time. Univariate procedures cannot provide insight on the former, as each variable is examined in isolation. As a result of this inability to consider the entire multivariate response space, the practice of following up a significant MANOVA result with individual ANOVAs does not provide insight to questions regarding multivariate differences (e.g., Huberty & Morris, 1989). Harris (2001) suggested that the use of MANOVA for between-group comparisons is more appropriate in the context of multiple dependent variables compared to the use of many individual univariate tests.

There is recognition that MANOVA may not be the best choice in all cases in which multiple outcome variables are of interest. The choice of the analytic procedure does rest on several factors including the data, research design, and research questions. For example, if the outcome variables are uncorrelated or have high positive correlations, then MANOVA may not be as effective as conducting separate univariate ANOVAs (Tabachnick & Fidell, 2007). In contrast, MANOVA can have greater power compared to the univariate methods when there is a moderate to strong negative correlation between the dependent variables (Tabachnick & Fidell, 2007). Additionally, power can depend on the relationship between dependent variables and the effect size (Cole, Maxwell, Arvey, & Salas, 1994). This study focuses on situations for which MANOVA may be most appropriate, based on recommendations from the works cited above, and considers the intercorrelations and effect sizes and how they relate to power of several test statistics as well as violations of assumptions, in order to highlight the

performance of these various test statistics associated with MANOVA, under different conditions.

To summarize the discussion heretofore, the decision regarding whether to select a univariate or multivariate comparison of between groups means must be made based on both statistical and substantive considerations. If the research questions are essentially multivariate in nature (e.g. Do the groups differ on the set of dependent variables?) then MANOVA is preferred to ANOVA (Stevens, 2001). In addition, when the dependent variables are at least moderately correlated, MANOVA will generally yield greater power compared to the univariate alternatives. Conversely, if the research questions are focused on the individual variables (e.g. Do the groups differ on Y1? Do the groups differ on Y2?), and/or if the dependent variables have little or no correlation or very strong positive correlations among them, then use of individual ANOVAs rather than MANOVA may be most appropriate (Stevens, 2001). In conclusion, the advantages of MANOVA, beyond Type I error control, can include (a) improving power for identifying group differences, (b) observing differences possibly missed in single ANOVAs (Huberty & Morris, 1989; Tabachnick & Fidell, 2007), and (c) understanding the outcome variables as a system rather than isolated measurements (Huberty & Morris, 1989). This study was conducted to examine performance of the several MANOVA test statistics in the case where multivariate questions are of primary interest and the multivariate procedure would be preferred.

Standard parametric multivariate means comparisons

In evaluating multivariate mean differences with MANOVA in the 2 group case, researchers test the null hypothesis of no group mean vector differences using Hotelling's T^2 statistic. Please see Johnson & Wichern (2002) for additional information on these multivariate test statistics. Hotelling's T^2 statistic which takes the form:

$$T^2 = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left[S\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{-1} \left(\bar{Y}_1 - \bar{Y}_2\right) \quad (1)$$

Where

$\bar{\boldsymbol{Y}}_1 =$ Mean vector for group 1

$\bar{\boldsymbol{Y}}_2 =$ Mean vector for group 2

$n_1 =$ Sample size for group 1

$n_2 =$ Sample size for group 2

$\boldsymbol{S} =$ Sample pooled covariance matrix; $\dfrac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$

$S_1 =$ Covariance matrix for group 1

$S_2 =$ Covariance matrix fro group 2

In this equation, the transpose (') operator is used to create sums of squared differences, in the context of matrices, and the inverse (-1) is used for matrix division. Hotelling's T^2 has been extended to accommodate the case of more than two groups with four different F approximation tests: Pillai's trace, (P) Wilk's lambda (Λ), Hotelling-Lawley Trace (H) and Roy's Greatest Root (R). These test statistics can be expressed as follows:

$$\Lambda = \frac{\lfloor W \rfloor}{\lfloor W+B \rfloor}$$

where (2)

$W =$ within group sum of squares and cross products matrix

$B =$ between group sum of squares and cross products matrix

$$P = tr\left[B(B+W)^{-1} \right] \quad (3)$$

$$H = tr\left[BW^{-1} \right] \quad (4)$$

$R =$ maximum eigenvalue of $W(B+W)^{-1}$ (5)

$|W| =$ Determinant of matrix W, where the determinant can be viewed as generalized or total variance of that matrix

Prior research regarding standard MANOVA test statistic performance

Accurate use and interpretation of these multivariate test statistics is dependent upon the assumptions of independent errors, multivariate normality, and homogeneity of group covariance matrices. When these assumptions are met, the tests perform similarly well with respect to controlling Type I error rates and maintaining appropriate statistical power, particularly in studies with relatively large sample sizes (e.g., Blair, Higgins, Karniski & Kromrey, 1994; Hopkins & Clay, 1963; Johnson & Wichern, 2002; Ramsey, 1982; Stevens, 2001). Several works cited in this review have informed multivariate researchers on how these statistics perform under various conditions. However, this work has primarily

been focused on the 2 group case. In addition, some of this work, particularly Ramsey, treated the data in a univariate fashion, rather than testing multivariate hypotheses about group means on several dependent variables simultaneously. Though this may be appropriate in some cases, many times where multivariate data are present, the hypothesis of interest concerns group differences on the set of means rather than on the individual means, in which case such univariate treatment of the data may be inappropriate (Huberty & Olejnik, 2006).

The work presented here focuses on the situation where researchers are interested in conducting multivariate means testing (rather than univariate), and is unique as (a) many test statistics are compared in a single simulation study, including a latent variable approach, and (b) the 3 group case is considered to ascertain whether the results from the 2 group case can generalize to the 3 group case, certainly a more complex but also perhaps more realistic condition. Many of these methods have been examined in simulation studies. However, the methods included here have not all been examined in a single study. Therefore, though it has been possible to describe how two or three of these statistics perform relative to one another, this study allows for the comparison of all of these methods under the same conditions.

Violations in assumptions of multivariate normality and homogeneity of covariance are often characteristic of social science research, and standard parametric MANOVA has limitations under such conditions (Blair et al., 1994; Everitt, 1979; Finch, 2005). Investigations of Type I error rates and power have suggested that these multivariate tests may not perform well when there are violations in assumptions of multivariate normality and equality of covariance matrices (e.g., Hakstian, Roed & Lind, 1979; Hopkins & Clay, 1963; Olson, 1974; Lee, 1971; Pillai & Jayachandran, 1967). Perhaps most notable is the performance of Hotelling T^2 in studies of unequal sample sizes when the assumptions of multivariate normality and particularly equality of covariance matrices has not been met. In such cases, the T^2 demonstrated diminished power as the degree of skewness of the response variables increased (Everitt, 1979). Furthermore, when the groups' covariance matrices were not homogeneous, the Type I error rate of the T^2 was inflated when the groups were not of equal size and the smaller group had the larger variances (Hakstian, Roed & Lind, 1979; Hopkins & Clay, 1963).

These results for T^2 are similar to those reported in studies of the performance of Pillai's Trace, Wilk's Lambda, Hotelling-Lawley's Trace and Roy's Greatest Root when there are violations in the assumption of equality of covariance matrices (Finch, 2005; Olson, 1974; Sheehan-Holt, 1998). In these

studies, when the smaller group had the larger variance the Type I error rates were inflated, whereas when the larger group had the larger elemental covariance elements, there was a reduction in power. Non-normality characterized by relatively severe skewness also resulted in a reduction of power (Everitt, 1979; Finch, 2005). Furthermore, when the assumptions were violated, Pillai's Trace was relatively more robust in terms of Type I error rate control compared to Wilk's Lambda and Hotelling-Lawley's Trace but exhibited somewhat lower power compared to these other tests. Not one of the common MANOVA statistics can be clearly identified as the single best test for use in all situations (Lee, 1971; Pillai & Jayachandran, 1967). The comparative effectiveness of these methods changed relative to specific features of the data. However, taken across a broad sweep of real data conditions, Λ, P and H all generally perform similarly, particularly when standard assumptions are met (Johnson & Wichern, 2002). In summary, the standard test statistics used with MANOVA are deleteriously affected by violations of the assumptions of normality and homogeneity of covariance matrices, particularly when samples are of unequal sizes.

Alternative test statistics to standard MANOVA when assumptions are violated

In response to these problems associated with assumption violations, a number of alternative test statistics have been investigated, particularly for use in the absence of multivariate normality and when group covariance matrices are not equal. The formulas for many of the basic versions of these statistics appear in Appendix A for the interested reader. Table 1 provides summary information across the different statistical tests to assist with organizing the information.

Brown and Forsythe (1974), James (1954), Johansen (1980), Yao (1965) and Nel and van der Merwe (1986) each outlined alternatives to the standard multivariate test statistic in the presence of unequal covariance matrices. Extensions of Hotelling's T^2, these parametric multivariate alternatives examine multiple outcomes between two groups, and have been extended for use with more than two groups. In the two groups case, the James (F_{JA}), Johansen (F_{JN}), Nel and van der Merwe (F_{NV}), and Yao (F_Y) statistics are based on the multivariate analog of the univariate *t*-test equation for unequal variances. $T^2_{unequal} = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} \left(\bar{Y}_1 - \bar{Y}_2\right)$ As with the univariate version of this statistic, the group variances (covariance matrices in the multivariate context) are

Table 1. General Conclusions based on the Literature of Test Statistics Examined for MANOVA Under Various Assumptions Conditions

	Assumptions	
Statistic	***Met***	***Not Met***
Standard (*P, H, L*)	Type I error rate controlled; Optimal power	Inflated Type I error for unequal covariance matrices and reduction of power for severely skewed data
F_{JA}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{JN}, F_{NV}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{Y}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{BF}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{K}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices but not to non-normal data.
TF_{J}, TF_{JN}	Comparable results to the standard test statistic.	For skewed and heavy tailed data, displayed higher power and better Type I error control than did F_{JN}.
TF_{NV}, TF_{Y}, TF_{BF}, TF_{K}	Comparable results to the standard test statistic.	For skewed and heavy tailed data, displayed higher power than did F_{K}.
Rank based test	Comparable Type I error rates to standard test but lower power.	For unequal covariance matrices, displayed better Type I error control though rates were still inflated.
SEM	Comparable Type I error and power rates to standard test for samples of 100 or greater.	Better Type I error control and higher power rates than standard tests for unequal covariance matrices

Note: T^2 = Hotelling's (1931); BF = Brown&Forsythe (1974), J = James (1954); JN = Johansen (1980),K= Kim(1992);NV= Nel & van der Merwe (1986),Y=Yao (1965), SEM = Structural Equation Modeling (Raykov, 2001), T with test = trimmed.

not pooled. The difference between F_{JA} and F_{JN} is in the way that they determine the critical value for assessing statistical significance. The F_{JA} statistics is simply $T^2_{unequal}$ (See Appendix A) with the critical value based on the χ^2 distribution adjusted by a complex term involving the traces of the covariance matrices for the

two groups. In contrast, the value for F_{JN} involves the conversion of $T^2_{unequal}$ to an *F* value, as seen in Appendix A.

The F_{NV} test statistic also is a transformed version of $T^2_{unequal}$ (see Appendix A) and compared to a critical *F* value. Krishnamoorthy and Xia (2006) presented a modified version of the degrees of freedom for F_{NV} labeling their statistic the Modified Nel and van der Merwe test (F_{MNV}). The test statistic remains the same, but the resulting value is compared to a critical *F* value with *p*, v_{KX} degrees of freedom, and the resulting test is affine invariant (results of the test are invariant under a linear transformation of the data). For a more thorough treatment of the calculation of v_{KX} the interested reader is encouraged to read Krishnamoorthy and Xia. Finally, among this set of statistics based upon the $T^2_{unequal}$ value is Yao's F_Y, which incorporates a different weighting scheme involving the determinant of the ratio of group covariance matrices (See Appendix A). Given that these previously described methods share a common root, namely $T^2_{unequal}$, they are discussed as a set of test statistics (i.e., Family 1). An examination of Appendix A reveals that although these statistics share a common root, they vary in terms of how they weight the groups' covariance matrices, and how degrees of freedom are calculated.

Of the alternatives to the standard T^2 described here, the Brown and Forsythe (F_{BF}) and the Kim (F_K) tests are not based on the $T^2_{unequal}$ statistic. The centerpiece of F_{BF} is T_{BF}, which differs from the $T^2_{unequal}$ statistic in terms of how the group covariance matrices are weighted, as can be seen in Appendix A. Essentially, where $T^2_{unequal}$ weights them by the inverse of sample size, T_{BF} uses the proportion of the total sample *not* in a specific group as the weight. Otherwise, T_{BF} is generally similar to $T^2_{unequal}$. The F_{BF} statistic is then compared to the critical value $F_{vBF1,vBF2}$. Kim's (F_K) statistic also is based on an alternative to $T^2_{unequal}$ and is compared with the $F_{m,vk}$ critical value. The calculation for F_K can be found in Appendix A. In general, it differs from both $T^2_{unequal}$ and T_{BF} in the way in which the group covariance matrices are weighted and combined. A review of Appendix A demonstrates that F_K relies on a more complex weighting system to combine these covariance matrices, using as a weight the determinant of their ratio (in the simplest two groups case) raised to the 1/(2*number of predictor variables) power. To simplify further discussion, and given their

similarity in terms of calculation, as mentioned previously F_{JN}, F_{NV}, F_{Y}, and F_{JA} have been organized into one family (Family 1) of statistics, and F_{BF} and F_{K} constitute a another family of test statistics (Family 2).

Prior research regarding alternative MANOVA test statistic performance

The test statistics in Families 1 and 2 have demonstrated relative robustness to the presence of unequal group covariance matrices (see Algina, Oshima, & Tang, 1991), which is reasonable given that their focus is on accounting for this condition by not relying on the pooled covariance matrix, *S*. Furthermore, the performance of these alternatives has proven to be superior to that of the standard Hotelling T^2 when data are multivariate normal but covariance matrices are unequal, both in terms of Type I error rates and power (Holloway & Dunn, 1967). However, these statistics are sensitive to non-normality in the form of moderate to severe skewness (Algina et al., 1991). Coombs, Algina, and Oltman (1996) investigated the Type I error rates of five multivariate generalizations of the Brown-Forsythe and Nel-van der Merwe tests and found that both F_{BF} and F_{NV} were able to maintain the nominal Type I error rate when heterogeneous group covariance matrices were present, but proved to be conservative when the ratio of total sample size to number of dependent variables was small. Christensen and Rencher (1997) observed increases in Type I error rates of F_{JA} and F_{Y}, particularly when the ratio of sample size to number of outcome variables was small. These authors recommended the use of F_{K} for cases in which the group covariance matrices were unequal. However, they acknowledged that this statistic was very conservative for cases in which the sample size to outcomes ratio was between 2 and 3. In a similar fashion, the F_{BF} and F_{NV} tests were shown to be conservative when the assumption of equal covariance matrices was violated and the sample size to outcome variables ratio was small (Coombs, Algina, and Olman, 1996). Additionally, Krishnamoorthy and Xia (2006) reported that F_{MNV} was able to maintain the nominal Type I error rate when group covariance matrices were unequal, as long as the response variables were distributed as multivariate normal. When the latter condition was not met, their test will likely not be appropriate as it relies on multivariate normality. Yanagihara and Yuan (2005) also examined many different versions of modified tests (e.g., *F* statistic, Bartlett correction, modified Bartlett correction) showing that the modified

Bartlett was comparable to the *F* statistic in many cases. This summary of work represents many studies that have examined various test statistics in the MANOVA framework to find a balance between Type I error and statistical power assist in the obtainment of an accurate statistical conclusion.

When the assumption of multivariate normality is violated these parametric MANOVA alternatives exhibit inflated Type I error rates, particularly with small sample sizes (Algina et al, 1991; Fouladi & Yockey, 2002; Wilcox, 1995). Thus, it appears that these alternative statistics are preferable to the standard multivariate test statistics when there are unequal group covariance matrices and the data are normally distributed. However, collectively they do not appear to be robust to violations of multivariate normality, yielding inflated Type I error rates.

Robust alternative test statistics for MANOVA

An alternative approach to the multivariate test statistics when there are violations of the normality assumption involves the use of trimmed means and Winsorized variance (Lix & Keselman, 2004). Statistical problems associated with nonnormality (e.g., Type I error inflation) in the univariate case can be ameliorated by using trimmed means and Winsorized variances in the construction of test statistics (e.g., Lix & Keselman, 2004; Keselman, Kowalchuk, & Lix, 1998; Wilcox, 1995). The use of the trimmed mean involves the removal of the most extreme data points of the response variable in each tail of the observed data distribution. The goal of such a statistic is to avoid the biasing of the mean estimate as a function of one or more outliers in the sample data. Wilcox (1995) recommended censoring 20% of the extreme observations at each tail of the distribution.

The appropriate measure of variation to accompany the trimmed mean is the Winsorized variance (Yuen, 1974). This estimate of variance is based on the Winsorized mean, which is calculated by replacing some portion (e.g., top and bottom 20%) of the most extreme scores in the sample data distribution with the next most extreme scores. The calculation for the Winsorized variance for variable *p* can be seen in Appendix A. As an example of trimming, consider the following set of 10 height measurements in inches: 58, 60, 69, 70, 70, 71, 71, 72, 73, 74. If the recommended 20% trimming were used, a total of 10 x 0.2, or 2, scores are removed. Thus the lower bound value (Y_L) is 60 and the upper bound value (Y_H) is 73, meaning that 58, 60, 73 and 74 are removed from each tail of the distribution, and thereby left out of the calculation of the trimmed mean, which in this case is 70.5. In contrast, the mean based on all 10 observations is 68.8. This is

how trimming was conducted for this study with SAS macros written by Lix and Keselman (2004). In other words, trimming and Winsorizing were conducted along each dimension individually, as described by Lix and Keselman. The Winsorized mean, which will be used in the calculation of the Winsorized variance, is based on 10 data points, with the lowest two values (58 and 60) replaced by 69, and the highest two values (73 and 74) replaced by 72. The value of $\overline{Y}_{wp} =$ 70.5, a 1.7 increase in the value used as the mean.

Lix and Keselman (2004) demonstrated how Winsorized variances and covariances can be applied to multivariate statistics in order to create a Winsorized covariance matrix. Note that the null hypothesis being tested when trimmed means are used involves only the part of the population of interest for which the trimmed mean is appropriate. Thus, the null hypothesis applies to population trimmed means. Given the trimmed means and Winsorized variances for a set of outcome variables, robust alternatives to the test statistics described above can be computed. Specifically, Lix and Keselman (2004) showed that both T^2 and $T^2_{unequal}$ can be calculated using the trimmed means and Winsorized covariance matrices. Likewise, the version of Hotelling's T^2 that does not use the pooled covariance matrix is available. See Appendix A. The robust test statistics will be organized into families using the same logic as described above for their non-trimmed versions; i.e. the trimmed versions reside under their home family (1 or 2).

Prior research regarding robust MANOVA test statistic performance

A number of the MANOVA test statistic alternatives described above based on trimmed means and Winsorized variances have been empirically compared (Wilcox, 1995). Wilcox focused on the case with 4 response variables, with a variety of data distributions, correlations among the response variables and sample sizes. Results showed that when the data were normally distributed, the standard and robust (trimmed) statistics exhibited comparable Type I error rates. However, for non-normal distributions (whether skewed or heavy tailed), the trimmed statistics F_{TK} and F_{TJN} were found to be preferable to their non-trimmed counterparts F_K and F_{JN} in terms of power, and overall, F_{TJN} demonstrated superior control over the Type I error rate for most of the simulated conditions. Beyond Wilcox's (1995) work, there is little empirical work comparing the performance of the robust alternatives to the other alternatives for multivariate mean comparisons when the group covariance matrices are not equal (Lix &

Keselman, 2004). It would appear, therefore, that an extensive evaluation of these methods under a variety of data conditions is warranted. Such work would inform the practitioner of which option may be optimal for use given data conditions. It also is noted that prior comparisons of these methods have been constrained to the two group case.

Rank based nonparametric test

Another alternative approach to dealing with violations of the standard MANOVA assumptions comes in the form of a rank based nonparametric test. A version of this test was first described by Puri and Sen (1971), and then further developed (Erdfelder, 1981; Katz & McSweeney, 1980). The statistic uses the ranks of the raw data as the dependent variables. Erdfelder's extension of this work involves the conversion of the Pillai's trace value obtained from conducting MANOVA using the ranks into the chi-square statistic $\chi^2 = (n-1)P$ (6), where P is Pillai's trace and n is the total sample size. The resulting value is compared with the χ^2 distribution with $k(p-1)$ degrees of freedom, where k is the number of groups for the independent variable and p is the number of response variables as described above. Thus, to compute this rank based nonparametric test, the researcher would first rank each of the dependent variables, and then conduct the MANOVA with the software package of choice, using the ranked dependent variables. The resulting value of P for the independent variable would then be converted using the equation described above. The rank based test represents a third family (Family 3) of statistics considered in this study.

Prior research regarding rank based MANOVA test statistic performance

There has been some empirical evaluation of the performance of the rank based approach, particularly as it compares to the common parametric statistics when the assumptions of normality and/or homogeneity of covariance matrices were violated. Ittenbach, Chayer, Bruininks, Thurlow, and Beirne-Smit (1993), for example, compared the rank based test with the standard MANOVA test statistics and reported somewhat higher power rates for the rank approach. However, Ittenbach and colleagues employed a real dataset for which the population distribution and equality status of the group covariance matrices was not known. Finch (2005) conducted a Monte Carlo simulation study comparing the rank based test statistic with Pillai's trace under a variety of conditions (e.g., normal and non-normal distributions, equal and unequal covariance matrices). When both

assumptions were met, Pillai's Trace and the nonparametric rank test each maintained Type I error rates near the nominal level, but the rank test exhibited lower power. When the assumption of normality was violated, both statistics maintained the nominal Type I error rate of 0.05, regardless of the type of distribution (double exponential, skewed normal, uniform), and had comparable power rates. In the presence of unequal covariance matrices, Finch noted that the rank based nonparametric tests resulted in lower Type I error rates compared to the parametric approach, though both methods had inflated values. Furthermore, as with standard multivariate statistics, the Type I error inflation when there were violations in covariance matrices was more pronounced when group sizes were unequal and the smaller group had the larger variances. Thus, the rank based alternative represents an improvement in the case of unequal covariance matrices, but may not be an ideal solution.

Structural equation models for MANOVA tests

Raykov (2001) suggested the use of structural equation modeling (SEM) as a potential alternative to MANOVA for testing the equality of group mean vectors, particularly when the assumption of equal covariance matrices is violated. He argued that because in the SEM framework covariance matrices can be allowed to differ, this approach might prove superior to the standard MANOVA when group covariances are heterogeneous. This may be an important property, given the aforementioned evidence that other MANOVA test statistics appear to have difficulty in both controlling Type I error and maintaining high power in the heterogeneous covariance case. The basic approach in this case is based on the standard confirmatory factor analysis (CFA) model (see Raykov, 2001), which takes the form:

$$x = \Lambda\xi + \delta$$

where

x = observed variable

ξ = vector of latent variables with covariance matrix Φ

Λ = factor loading matrix

δ = error term

(7)

In most applications of CFA, each latent variable is associated with multiple observed variables. However, in this case each observed dependent variable in the

MANOVA context is related to its own unique latent variable, due to the following strictures:

$$\Lambda = I_p \text{ and } \Theta = 0_{pxp} \tag{8}$$

Here I_p is the identity matrix and Θ is the covariance matrix for δ. In this special case, the covariance matrix for error is comprised of zero elements. These special restrictions, taken together with the CFA model imply that each latent variable is equal to one of the observed variables (Raykov, 2001) and that the latent variable covariance matrix is identical to that of the observed variables. In order to test the null hypothesis of equality of group mean vectors for the response variables, two further assumptions must be made (Raykov, 2001):

$$\begin{aligned} &(1)\ E(\xi) = E(\mu) \\ &(2)\ E(\delta) = 0 \end{aligned} \tag{9}$$

These additional restrictions to the model make the comparison of latent means equivalent to a comparison of observed means. The researcher can then test the null hypothesis of no group difference on the vector of observed dependent variable means by fitting two CFA models, one in which the response variable means are constrained to be equal across groups and the other in which they are allowed to vary. Then, the test of the null hypothesis of group difference on the responses is the difference in the χ^2 fit statistics: $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}}$ (10). Allowing the group means to differ results in a saturated CFA model so that the value of $\chi^2_{\text{Constrained}}$ will be 0. Therefore, the test of the null hypothesis of group differences across the vector of dependent variable means is equivalent to $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}} = \chi^2_{\text{Constrained}} - 0 = \chi^2_{\text{Constrained}}$ (11; Raykov, 2001).

As noted above, the primary advantage of using the SEM approach to compare group mean vectors is that covariance matrices can be allowed to vary across groups (Raykov, 2001). In this way, the assumption of covariance matrix equality which underlies standard MANOVA and which has been shown in prior research to be important for other statistics for testing multivariate mean equality, is no longer a requirement. When the assumption of normality is violated, the standard χ^2 statistic used with ML estimation in SEM may not perform well (Yu & Muthén, 2002). Therefore, an adjusted version of this test statistic is appropriate when the dependent variables are not normally distributed. This alternative, developed by Satorra and Bentler (1994), was designed to correct for

multivariate kurtosis, and has been shown to be robust to departures from multivariate normality (Curran, West, & Finch, 1996).

Given that the MANOVA test statistics are not as accurate as desired under violations of assumptions, alternative methods need to be explored to test the same hypotheses (Raykov, 2001) yet such evaluations have not occurred in sufficient number. As described above, prior simulation research examining alternatives to the standard MANOVA approach for testing multivariate mean differences (e.g., rank based and exact tests) has generally found that assumption violations, particularly that of homogeneity of covariance matrices, result in Type I error inflation similar to, if not as severe as, that reported for MANOVA (e.g., Finch, 2005; Ittenbach, Chayer, Bruininks, Thurlow, & Beirne-Smith, 1993). By contrast, very little empirical research has been conducted to evaluate the effectiveness of this fairly new SEM based approach for testing the null hypothesis of multivariate mean equality. One such effort (Finch & French, 2008) found that in the absence of assumption violations, the Satorra-Bentler corrected χ^2 test and Pillai's trace had comparable Type I error rates and power for total samples of 100 or more with normally distributed dependent variables. For smaller samples, the SEM based approach did have elevated Type I error rates (e.g., 0.09 for *N* of 30) when both assumptions of normality and homogeneity of covariance matrices were met. When the assumption of equal covariance matrices was violated and the smaller group had the larger elements, the SEM based approach had lower Type I error rates compared to the standard approach. When the larger group had the larger elements, both SEM and the standard approach had Type I error rates at or below the nominal level, but the SEM method had much higher power. Thus, it appeared that the SEM approach might be preferred. However, there is a need to examine the large number of viable MANOVA test statistics reviewed here under the same conditions to begin to inform the field as to which approach is preferred under different conditions. Additionally, little, if any prior work has examined the performance of this new SEM approach to MANOVA testing as well as with more than two groups. The SEM approach to testing hypotheses about multivariate mean differences represents a fourth family (Family 4) of test statistics investigated in this study.

Goals of the study

The first goal of this study was to review the various MANOVA test statistics to inform the reader of the 16 choices that are currently available for comparing multivariate means across groups. Table 1 provides summary information across

these 16 tests to aid understanding of performance from separate past evaluations. The second goal was to conduct a simulation study comparing the performance of the 16 methods across a variety of conditions designed to mirror those encountered in practice, in order to assess their Type I error and power rates. This Monte Carlo study is anticipated to provide information on performance of these tests to aid the researcher in selecting the test that appears to work well given the specific data at hand and corresponding assumptions that are or are not met. The literature review led to several predictions for comparing test statistics noting that it is impractical to make predictions for all combinations investigated. First, it was expected that when the data were normally distributed and group covariance matrices were homogeneous, all methods would have comparable Type I error and power rates. Second, Families 1, 2, 3 and 4 were expected to have, on average, lower Type I error and higher power compared to the standard MANOVA test statistic, when covariance matrices were heterogeneous. Third, given the advantages of latent variable modeling it was expected that SEM would have the lowest Type I error and highest power, across conditions, with the exception of for small sample sizes, where accurate parameter estimation would likely be a problem. Fourth and last, trimmed versions in Family 1 were expected to have the lowest Type I error and highest power in heavily skewed distribution conditions.

A number of studies have previously conducted investigations of a few of these methods, but no study has simultaneously compared all of the techniques under a common set of conditions. In addition to all comparisons under similar conditions, this work adds to the literature by providing information on the use of SEM under these conditions and behavior of all statistics studied for the 2 and 3 group case. The former is rarely included in such comparisons and no evaluation has investigated performance of all four test families in one simulation under the same conditions. Thus, the present work seeks to extend the literature by providing a full examination of methods for comparing multivariate group means when standard assumptions are not met. A total of seven factors were manipulated which allowed for the examination of 12,076 conditions to assist with meeting the second goal of the study.

Methods

This Monte Carlo study manipulated seven factors in a completely crossed design with 5000 replications per combination of conditions using a SAS program (SAS version 9.1, 2004) written by the authors. Manipulated factors included sample size, group size ratio, covariance matrix homogeneity/heterogeneity, distribution

of the dependent variables, group mean differences, correlations among the dependent variables, and the number of dependent variables. All of the statistical methods were conducted using SAS, with the exception of SEM, which was carried out with M*plus* version 5.1 (Muthén & Muthén, 2008). A number of the alternative and robust methods were conducted with a macro described by Lix and Keselman (2004). The two outcome variables of interest were the Type I error rate (rejecting the null hypothesis of no multivariate mean difference when, actually, no differences were simulated) and power (correctly rejecting the null hypothesis of no multivariate mean differences). To assess which of the manipulated factors, or combinations of them, had a significant influence on the dependent variables, an ANOVA was conducted for each outcome, per recommendations for simulation research (Paxton, Curran, Bollen, Kirby, & Chen, 2001). The dependent variable in each ANOVA was one of the outcomes (i.e., Type I error rate or power) taken across replications for each combination of conditions. The independent variables were the manipulated factors and their interactions. In addition, the ω^2 effect size was calculated to describe the relative magnitude of the statistically significant effects. Given the scope of the simulation, discussion is limited only to those effects that most influenced the Type I error and power rates, which are defined as those effects that were both statistically significant ($\alpha = 0.05$) and had an ω^2 of 0.10 or greater.

Statistical methods

Because it has been demonstrated as more robust with respect to Type I error control when standard assumptions are violated (Olson, 1974), *Pillai's Trace* (*P*) was selected for use as the standard MANOVA test statistic for this study, and will be referred to as such throughout the remainder of the manuscript, although it is acknowledged that other test statistics such as Wilks' Lambda, are also frequently used in practice. However, note that with the two groups case the results across the standard tests will be identical, and equal to Hotelling's T^2. The other statistical tests included the rank based method, James (JA), Hotelling's T^2 for unequal covariance matrices (H), Brown-Forsythe (BF), Johansen (JO), Kim (K), Nel van der Merwe (NV), Yao (Y), Raykov (SEM), and the trimmed versions of the robust methods, TJA, TH, TBF, TJO, TK, TNV, and TY. Consistent with the recommendation of Lix and Keselman (2004), 20% symmetric trimming of the data was employed.

Manipulated Factors

Total sample size

Seven total sample size (across groups) conditions were examined for the two groups case: 20, 30, 45, 60, 90, 100, and 150. For the three groups case, the following total sample size conditions were examined: 30, 40, 45, 50, 60, 75, 90, 120, 150, 200, and 250. In the three groups, equal sample size condition for N=40, 50, 200, and 250, the data were simulated so that one group had either one more or one fewer observations than did the others. For example, in the N=40 case, two of the groups were simulated with 16 individuals, whereas the other was simulated with 17. Similarly, in the N=250 condition, two of the groups were simulated with 83 individuals, whereas the other was simulated with 84. The same approach was used with 50 and 200. These values are in accord with previous simulation research with MANOVA and SEM approaches to multivariate comparisons, (e.g., Christensen & Rencher, 1997; Finch, 2005; Hancock, Lawrence & Nevitt, 2001; Hussein & Carriere, 2005; Wilcox, 1995). This range of values was selected to reflect conditions that applied social science researchers are very likely to encounter.

Number of Groups

Two conditions were simulated for number of groups: 2 and 3 groups. Much of the previous work comparing performance in the MANOVA situation has been conducted on 2 groups with several variables (e.g. Christensen & Rencher, 1997; Finch, 2005). A significant addition of this work to the literature is to evaluate the behavior of these tests with 3 groups. Two groups were included to aid the comparison to prior work.

Group size ratio

Three group size ratio conditions were used: (a) groups were equal, (b) group 1 was half the size of group 2, and (c) group 1 was twice the size of group 2. In the three group case, for condition (b) groups 1 and 2 were half the size of group 3, and for condition (c) groups 1 and 2 were twice the size of group 3. Thus, for example, in the n=60 case, there were 30 simulees per group in condition a, 20 in group 1 and 40 in group 2 in condition b, and 40 in group 1 and 20 in group 2 in condition c. The combination of unequal group sizes with unequal group covariance matrices has been shown to influence both Type I error and power

rates (Sheehan-Holt, 1998; Stevens, 2001; Hakstian, Roed & Lind, 1979) and these particular ratios employed have been used in prior studies (e.g., Christensen & Rencher, 1997, Hakstian et al.,1979). As noted above, when the smaller group has the larger covariance matrix elements the Type I error rate will be inflated; when the larger group has the larger elements power will be diminished.

Covariance matrix homogeneity/heterogeneity

The group covariance elements were manipulated in three ways: (a) equal, (b) one group had elements 5 times as large as the others, and (c) one group had elements 10 times as large as the others. Equality of the covariance matrices across groups is a vital assumption for the test statistics associated with MANOVA, and differences in these matrices can influence the performance of these tests (Fouladi & Yockey, 2002; Sheehan-Holt, 1998; Korin, 1972). Two unequal covariance conditions were simulated (a) the larger group had the larger elemental values and (b) the smaller group had the larger elemental values.

Distribution of the dependent variables

Normality of the dependent variables is another assumption of the standard statistical tests used in MANOVA. The Type I error rate of the common MANOVA tests may suffer from some inflation when the distribution of the dependent variables have large kurtosis (Olson, 1974). Therefore, in the current research the dependent variables were simulated under one of four distributional conditions: (a) normal (skewness=0, kurtosis=0), (b) beta (skewness = -0.82, kurtosis = 0.28), (c) lognormal (skewness = 6.18, kurtosis = 110.93), and (d) exponential (skewness =2, kurtosis = 6). These reflect conditions used in similar work (Algina et al., 1991). The non-normal data were simulated using a methodology described by Headrick and Sawilowsky (1999) to achieve the desired levels of skewness and kurtosis while maintaining the target correlations among the dependent variables. These distributions were selected to provide insights into the performance of the methods studied here under a variety of cases.

Group means differences

Differences in group means were simulated using values of Cohen's (1988) *d* univariate effect sizes. This metric was selected because it allowed for a straightforward manipulation of this important variable and matches the methodology (though not the values) used in prior simulation research of MANOVA (Blair et al., 1994; Finch, 2005). The effect size of 0 allowed for the evaluation of the Type I error. The other values corresponded to group separation

at small (0,2), medium (0.5), and large (0.8) levels. The univariate Cohen's *d* (i.e., $mean_{group1} - mean_{group2} / SD_{pooled}$) effect size was selected for use in this study because there are generally agreed upon guidelines for its interpretation (Kim & Olejnik, 2004). In contrast, though there do exist multivariate effect size values, there is not a single such statistic that is considered the standard, nor is there any sort of agreement regarding what constitutes a small, medium, or large effect. Thus, in order to provide a useful context to researchers regarding the performance of the various methods described here, Cohen's *d* was used.

Correlation among the dependent variables

The data were simulated under three conditions for correlation among the dependent variables, including no correlation (0.0), small (0.2) and large (0.8). These values were selected to represent the case where variables were orthogonal (0.0), where the correlation was small to moderate (0.2) and where the variables were highly correlated (0.8). Conditions are consistent with prior research (e.g., Finch, 2005; Wilcox, 1995) to aid comparability.

Number of dependent variables

Two levels were employed: 2 and 4 dependent variables, consistent with prior studies (e.g., Fouladi & Yockey, 2002; Wilcox, 1995) and representative of realistic numbers of response variables seen in practice (e.g., Dumas et al., 1999; Krull, Kirk, Prusick, & French, 2010) while maintaining a manageable set of simulation conditions for the current study.

Results

Two groups versus three groups

Results for two and three group cases generally followed similar patterns in terms of how the methods compared relative to one another, with a couple of exceptions. Thus, to keep discussion of the results as brief as possible, only results for the three group condition are presented. However, prior to presenting these, note that the few cases where the two group condition results diverged from those for three group condition. In general, Type I error rates did not differ between the two number of groups conditions, but power was higher in the three group case compared to the two group case. In terms of relative comparison of the methods, with two groups the rank based approach had among the lowest power values. When three groups were present, the rank based approach had comparable power

to the other approaches, as is presented below. Outside of these differences, the results for the two group case were comparable to those for three group case, which are presented below. The two group case results are available from the authors upon request.

The results for three group case are organized into two sections: (a) Type I error and power rates based on the variance homogeneity condition, and (b) Type I and power rates by the distribution of the response variables. In each case, a repeated measures ANOVA was employed to identify the significant main effects and interactions of the manipulated factors in terms of Type I error and power, where the repeated measures variable was the MANOVA test statistic. The ANOVA models had as the dependent variable the Type I error or power rates across the 5000 replications per combination of conditions. The independent variables were type of test statistic (within replication), correlation among the dependent variables, number of dependent variables, sample size ratio, variance ratio, sample size and in the case of power, and effect size. The assumptions of normality and sphericity were assessed and found to have been met. Sphericity was assessed using Mauchley's test of Sphericity in conjunction with the ε statistic, which takes the value of 1 in the population when the covariance matrix satisfies sphericity (Warner, 2008). In the case of each set of repeated measures ANOVA results below, Mauchly's test was not statistically significant with α = 0.15, as recommended in Kirk (1995). In addition, across the repeated measures analyses described below, the value of ε ranged between 0.901 and 0.974. Finally, an examination of the Greenhouse-Geisser conservative *F*-test and MANOVA test results, both of which have been suggested for use when sphericity is violated, revealed the same main effects and interactions as significant and non-significant when compared with the unadjusted test. Therefore, given the general finding that sphericity was present, coupled with the similarity in results for the unadjusted and Greenhouse-Geisser adjusted test, it may be concluded that sphericity (or lack thereof) was not problematic in this case.

Normality was assessed first by an examination of QQ-plots for the individual outcome variables, and all were found to conform reasonably closely to the line for the normal distribution. In addition, multivariate normality was tested for across repeated measurements (rejection rates for each test statistic) for each of the models described below using Mardia's test (Mardia, 1970), and found none of them to be statistically significant. Taken together, the QQ-plot and Mardia's test results satisfy the assumption of normality for repeated measures models as described in Warner (2008). The models were fully factorial with all main effects and interactions included. As mentioned previously, in order to focus

on only the most important of the manipulated factors, discussion will be limited to those significant (α = 0.05) terms in the ANOVA that had an effect size (ω^2) greater than 0.10. This value was selected because it indicates that the main effect or interaction term accounted for at least 10% of the variation in rejection rates. By doing so, it is possible to avoid discussing a large number of statistically significant effects that actually accounted for a small amount of variance, which was a concern given the large number of replications for each combination of conditions. Full results tables are available by contacting the authors.

Covariance Homogeneity: Type I error rate

The ANOVA identified the interaction of test statistic by sample size ratio by covariance ratio as the highest order significant term ($p < 0.01$, $\omega^2 = 0.527$). The interaction of test statistic by number of dependent variables was also significant ($p < 0.01$, $\omega^2 = 0.381$). No other term was statistically significant with an effect size value greater than 0.10.

Table 2 contains the Type I error rates by test statistic, sample size ratio, and covariance ratio for normally distributed data. When the groups' covariances were equal, the Type I error rate for all of the statistics examined here were below 0.06, except for H, TH, and the rank approach across group ratio conditions, and for BF in the sample size ratio 2/1 condition. When the group covariances were not equal but the sample size ratio was equal, the Type I error rate of the *P* test was inflated above the nominal 0.05 level. Several of the alternative statistics, including the rank based approach, H, TH, and BF had inflated Type I error rates in the unequal covariance, equal sample size condition as well. In contrast, the Type I error rates for JA, JO, K, NV, Y, all members of Family 1 (except for K), and SEM did not have inflated error rates associated with inequality in group covariances when sample sizes were equal. To further investigate these effects, several interaction contrasts were employed, using Scheffé's correction (Scheffé, 1953) to control the overall Type I error rate (α = 0.05) and allow for such post hoc investigations. Based on these contrasts, it was found that the rank and H statistics yielded significantly inflated Type I error rates as the degree of covariance inequality increased, whereas the rates of the other methods did not change significantly.

Table 2. Type I Error Rate by Test Statistic, Sample Size Ratio, and Group Covariance Ratio: Normally Distributed Data

Sample Size Ratio	Statistic	Covariance ratio: 1/1	Covariance ratio: 5/1	Covariance ratio: 10/1
Equal	*Standard*	0.050	0.060	0.064
	Ranks	0.060	0.072	0.082
	JA | *TJA*	0.043 | 0.031	0.049 | 0.033	0.052 | 0.040
	H | *TH*	0.070 | 0.070	0.089 | 0.093	0.102 | 0.117
	BF | *TBF*	0.053 | 0.038	0.071 | 0.046	0.072 | 0.046
	JO | *TJO*	0.051 | 0.042	0.056 | 0.044	0.056 | 0.050
	K | *TK*	0.047 | 0.034	0.040 | 0.027	0.035 | 0.024
	NV | *TNV*	0.048 | 0.034	0.047 | 0.029	0.047 | 0.028
	Y | *TY*	0.048 | 0.035	0.055 | 0.040	0.059 | 0.044
	SEM	0.057	0.054	0.058
1/2*	*Standard*	0.050	0.007	0.004
	Ranks	0.061	0.023	0.019
	JA | *TJA*	0.044 | 0.033	0.048 | 0.042	0.046 | 0.040
	H | *TH*	0.061 | 0.061	0.031 | 0.034	0.024 | 0.018
	BF | *TBF*	0.052 | 0.043	0.064 | 0.075	0.067 | 0.081
	JO | *TJO*	0.051 | 0.045	0.051 | 0.041	0.046 | 0.040
	K | *TK*	0.051 | 0.042	0.046 | 0.039	0.042 | 0.036
	NV | *TNV*	0.047 | 0.035	0.047 | 0.042	0.044 | 0.041
	Y | *TY*	0.053 | 0.046	0.048 | 0.043	0.049 | 0.039
	SEM	0.053	0.055	0.061
2/1**	*Standard*	0.049	0.092	0.109
	Ranks	0.061	0.086	0.103
	JA | *TJA*	0.042 | 0.033	0.047 | 0.037	0.050 | 0.041
	H | *TH*	0.068 | 0.065	0.122 | 0.121	0.157 | 0.159
	BF | *TBF*	0.064 | 0.052	0.069 | 0.050	0.070 | 0.049
	JO | *TJO*	0.053 | 0.048	0.055 | 0.046	0.055 | 0.048
	K | *TK*	0.050 | 0.041	0.041 | 0.033	0.037 | 0.029
	NV | *TNV*	0.044 | 0.039	0.047 | 0.033	0.048 | 0.033
	Y | *TY*	0.058 | 0.047	0.055 | 0.047	0.055 | 0.046
	SEM	0.049	0.055	0.053

*Sample size ratio of 1/2 couples larger variance with larger group size in the unequal variance condition.
**Sample size ratio of 2/1 couples larger variance with smaller group size in the unequal variance condition.

Based on interaction contrasts using Scheffé's corrected critical value, when the larger group had the larger covariance (sample size ratio 1/2), the *P*, rank based statistic and H displayed significant declines in Type I error rates concomitant with increases in groups' covariance matrix inequality. As the group

covariances became more unequal, however, TBF had a significant increase in the Type I error rate. As seen with an equal sample size ratio, members of Family 1 and K generally demonstrated consistent Type I error rates, which were just below the nominal value of 0.05. The Scheffé corrected contrasts did not find any significant change in the error rates of the SEM method, though for the covariance ratio of 10/1 with the 1/2 sample size ratio, the rate was just above 0.06. When the smaller group had the larger covariance (sample size ratio 2/1), the standard, rank based, H, and TH approaches all showed a significant increase in the Type I error rate with increasing divergence in group covariance matrices. Family 1 and SEM maintained Type I error rates near the nominal 0.05 value, whereas K actually had a slight decline in the error rate as the covariance matrices became more unequal. Across all conditions simulated here, the trimmed versions of the test statistics had slightly lower Type I error rates compared to the untrimmed alternatives (except for TH in the covariance ratio 10/1, sample size ratio 2/1 case), though in most cases these differences were less than 0.01.

Table 3. Type I Error Rate by Test Statistic and Number of Dependent Variables: Normally Distributed Data

	Number of dependent variables	
Statistic	**2**	**4**
Standard	0.069	0.087
Ranks	0.065	0.079
JA | *TJA*	0.043 | 0.042	0.041 | 0.039
H | *TH*	0.072 | 0.074	0.074 | 0.078
BF | *TBF*	0.062 | 0.052	0.064 | 0.049
JO | *TJO*	0.049 | 0.044	0.051 | 0.042
K | *TK*	0.048 | 0.041	0.043 | 0.040
NV | *TNV*	0.050 | 0.046	0.052 | 0.039
Y | *TY*	0.046 | 0.046	0.053 | 0.038
SEM	0.057	0.051

Table 3 displays the Type I error rate for statistical test by number of dependent variables for normally distributed data. The error rates for the standard and rank based approaches were significantly greater for 4 variables compared to 2 variables. The Type I error rates for the rest of the test statistics were essentially the same for 2 and 4 dependent variables. In addition to the standard and rank approaches, H, TH, and BF all had error rates in excess of 0.06; the other methods

had rates closer to the nominal 0.05. Because there were not significant results for the correlation among the dependent variables and the sample size, they are not discussed.

Covariance Homogeneity: Power

Repeated measures ANOVA was used to identify the manipulated terms that were significantly related to power rates across replications, using the same model used with Type I error rates. The interaction of the test statistic by sample size ratio by covariance ratio was the highest order significant term ($p < 0.01$, $\omega^2 = 0.149$), as were the main effects of effect size ($p < 0.01$, $\omega^2 = 0.811$), correlation among the dependent variables ($p < 0.01$, $\omega^2 = 0.360$) and total sample size ($p < 0.01$, $\omega^2 = 0.781$). No other terms in the ANOVA were statistically significant with an effect size greater than 0.10.

Table 4 contains power by test statistic, sample size ratio and group covariance ratio. Power values for those conditions for which the Type I error rate was greater than 0.075 (from Table 2) are in bold, and should be interpreted with extreme caution. These values are included for completeness in results presentation. When the groups were of equal size, SEM, followed by the *P* statistic had the highest power rates among those for which the Type I error rates were not inflated (non-bolded values). For all of the methods studied here, power declined as the covariance matrix inequality increased when the larger group had the larger variance and when the smaller group had the larger variance. In addition, the power for the trimmed statistics was uniformly lower than that of the non-trimmed versions in this sample. Power for the rank based approach was comparable to that of the standard in the covariance 1/1 and 5/1 cases, but could not be interpreted for 10/1 due to Type I error inflation.

When the group sizes were unequal but the covariance matrices were equal, SEM had the highest power rates, followed by the standard, and rank based approaches, all of which had significantly higher power than the other methods studied here. When the larger group had the larger covariance (sample size 1/2 condition), power for all methods declined significantly with increases in variance heterogeneity, though the pattern of SEM, followed by standard and rank methods with highest power rates held. When the smaller group had the larger covariance (sample size 2/1 condition), a situation that resulted in inflated Type I error rates for several methods, the highest power rates among those that had Type I error rates lower than 0.075 belonged to SEM, followed by Family 1, K, BF, and TBF. For all of the methods power rates declined significantly as the degree of

covariance matrix inequality increased. Note that in this condition, the Type I error rates for the standard, rank based, and H approaches were inflated.

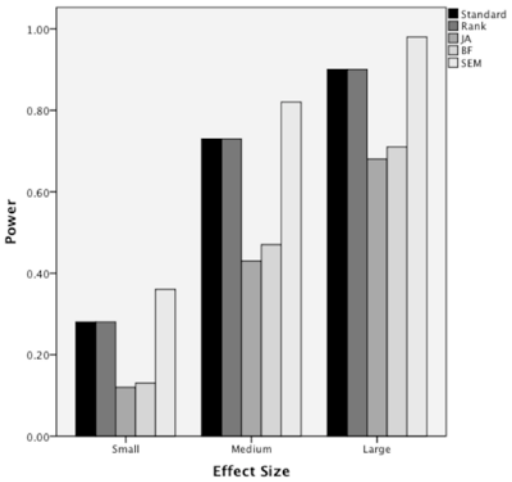
Table 4. Power by Test Statistic, Sample Size Ratio, and Group Covariance Ratio: Normally Distributed Data

Sample Size Ratio	Statistic	Covariance ratio: 1/1	Covariance ratio: 5/1	Covariance ratio: 10/1
Equal	Standard	0.695	0.44	0.309
	Ranks	0.684	0.464	**0.353**
	JA | TJA	0.470 | 0.394	0.256 | 0.203	0.170 | 0.131
	H | TH	0.530 | 0.496	**0.330 | 0.309**	**0.248 | 0.241**
	BF | TBF	0.495 | 0.421	0.302 | 0.230	0.209 | 0.148
	JO | TJO	0.490 | 0.432	0.268 | 0.223	0.178 | 0.145
	K | TK	0.480 | 0.402	0.240 | 0.190	0.142 | 0.102
	NV | TNV	0.483 | 0.405	0.253 | 0.189	0.162 | 0.112
	Y | TY	0.481 | 0.404	0.266 | 0.210	0.177 | 0.135
	SEM	0.738	0.489	0.357
1/2*	Standard	0.764	0.538	0.413
	Ranks	0.758	0.55	**0.435**
	JA | TJA	0.537 | 0.463	0.319 | 0.267	0.222 | 0.184
	H | TH	0.587 | 0.558	**0.389 | 0.369**	**0.312 | 0.300**
	BF | TBF	0.558 | 0.500	0.363 **| 0.300**	0.261 **| 0.206**
	JO | TJO	0.558 | 0.506	0.332 | 0.288	0.230 | 0.194
	K | TK	0.551 | 0.491	0.310 | 0.259	0.201 | 0.159
	NV | TNV	0.545 | 0.468	0.321 | 0.261	0.220 | 0.171
	Y | TY	0.557 | 0.499	0.332 | 0.283	0.229 | 0.188
	SEM	0.802	0.591	0.472
2/1**	Standard	0.741	**0.705**	**0.685**
	Ranks	0.734	**0.721**	**0.69**
	JA | TJA	0.514 | 0.440	0.498 | 0.403	0.377 | 0.334
	H | TH	0.568 | 0.537	**0.536 | 0.511**	**0.436 | 0.367**
	BF | TBF	0.537 | 0.473	0.492 | 0.397	0.381 | 0.326
	JO | TJO	0.535 | 0.482	0.488 | 0.401	0.379 | 0.319
	K | TK	0.527 | 0.461	0.487 | 0.410	0.384 | 0.322
	NV | TNV	0.525 | 0.447	0.500 | 0.389	0.380 | 0.343
	Y | TY	0.532 | 0.467	0.519 | 0.402	0.399 | 0.338
	SEM	0.811	0.732	0.691

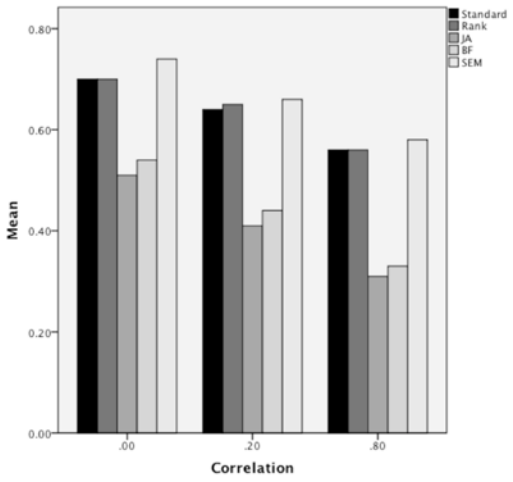
Note: Bold indicates when power values for these conditions were associated with Type I error rates greater than 0.075

*Sample size ratio of 1/2 couples larger variance with larger group size in the unequal variance condition.

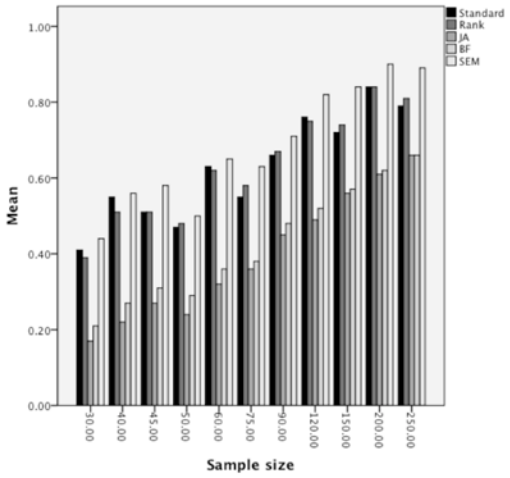
**Sample size ratio of 2/1 couples larger variance with smaller group size in the unequal variance condition.



Power by effect size



Power by correlation among dependent variables



Power by sample size

Figure 1. Three panels of power rates for five MANOVA statistics by effect size, correlation among dependent variables, and sample size.

Figure 1 displays power by the main effects of effect size, correlation among the dependent variables and total sample size, in three panels. For clarity of presentation only selected testing methods were included, as they are representative of others studied. Specifically, JA was selected to represent Family 1 (except H) and K, all of which had very similar rates, though BF displayed similar power to H under these conditions. The trimmed versions of these statistics had power rates that were similar to the untrimmed versions in terms of their pattern relative to one another and had slightly lower power values (though not significantly lower) than the untrimmed statistics. For all of the methods,

power increased significantly with increases in effect size and sample size, and declined with increases in the correlations among the dependent variables. These patterns were consistent across the methods studied here.

Distribution: Type I error rate

As with the covariance homogeneity data, a fully factorial repeated measures ANOVA was used to identify significant main effects and interactions of the manipulated variables that were related to the Type I error rates under differing distribution conditions. The highest order term that was identified as statistically significant with ω^2 greater than 0.10 was the interaction of type of test statistic (method) by number of dependent variables by sample size ($p < 0.01$, $\omega^2 = 0.624$). In addition, the distribution of the dependent variables was a significant main effect ($p = 0.034$, $\omega^2 = 0.063$). Although its ω^2 value did not meet the 0.10 threshold used to identify terms for further consideration, it will be discussed briefly because the distribution of the response was of primary interest in this study. No other term was both statistically significant in the ANOVA and had ω^2 greater than 0.10.

Table 5. Type I Error Rate by Test Statistic and Distribution of the Dependent Variables.

	Distribution			
Statistic	**Normal**	**Beta**	**Lognormal**	**Exponential**
Standard	0.05	0.05	0.05	0.05
Ranks	0.079	0.06	0.061	0.06
JA | TJA	0.047 | 0.036	0.044 | 0.032	0.044 | 0.033	0.044 | 0.032
H | TH	0.104 | 0.106	0.064 | 0.065	0.064 | 0.065	0.064 | 0.064
BF | TBF	0.064 | 0.046	0.052 | 0.042	0.052 | 0.042	0.052 | 0.041
JO | TJO	0.054 | 0.046	0.051 | 0.044	0.051 | 0.045	0.051 | 0.044
K | TK	0.042 | 0.032	0.049 | 0.039	0.048 | 0.040	0.048 | 0.039
NV | TNV	0.047 | 0.032	0.047 | 0.035	0.047 | 0.036	0.047 | 0.035
Y | TY	0.054 | 0.044	0.052 | 0.043	0.051 | 0.043	0.051 | 0.042
SEM	0.055	0.082	0.084	0.082

Table 5 contains the Type I error rate for the test statistics by the distribution of the dependent variables. These results demonstrate that the *P* test statistic was robust to the distribution of the dependent variables, maintaining the nominal (0.05) Type I error rate across the four distributions. With the exception of the

rank based approach, BF, H, and TH in the normal case, and ranks, H/TH, and SEM in the nonnormal conditions, which had elevated rates, the tests displayed Type I error at the nominal level of 0.05.

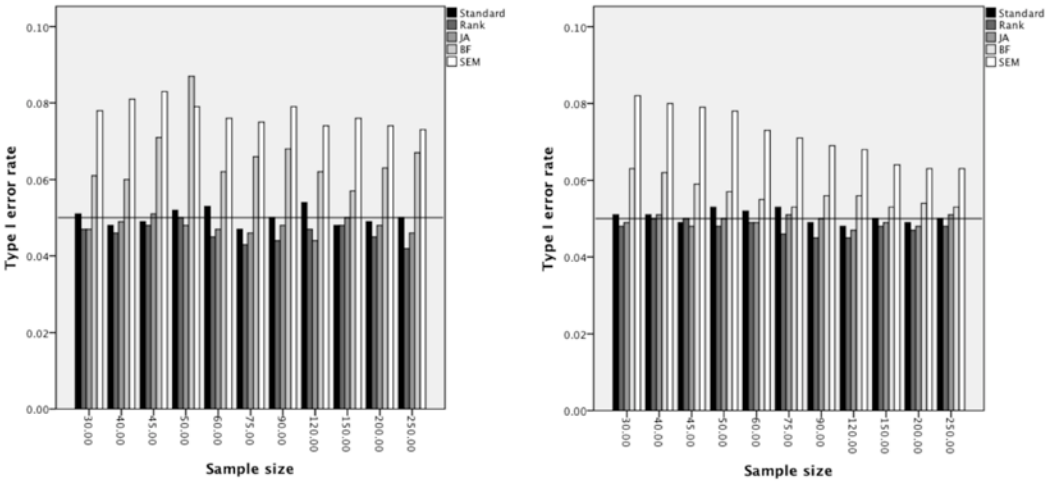


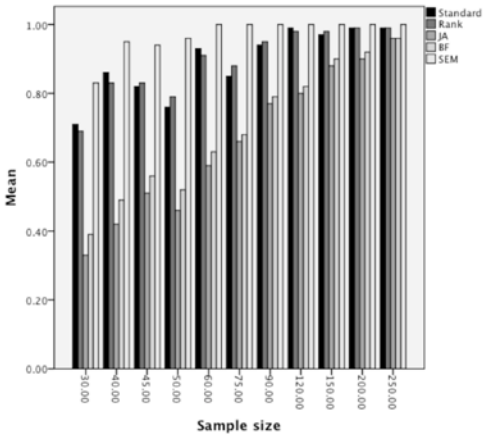
Figure 2. Two panels of Type I error rates for five MANOVA tests by sample size and number of dependent variables, across distribution of the dependent variables.

Figure 2 contains two panels showing the Type I error rates for the methods by the number of dependent variables and the sample size, across distribution conditions. In order to simplify presentation of the results, only the selected methods described were examined, which are representative of other several others that performed extremely similarly. An examination of Figure 2, which has a reference line at the nominal α rate of 0.05, reveals that when there were 2 dependent variables, BF and SEM consistently had elevated Type I error rates. The other methods largely maintained the nominal rate across sample sizes, although the standard statistic did have slightly rates slightly above the 0.05 line (though not as high as 0.06) at N=60 and 120. With 4 dependent variables the standard, rank, and JA methods exhibited Type I error rates near or just below the 0.05 level, except for the standard statistic with samples of 50, 60, and 75, with rates slightly above the nominal rate but not breaking 0.06. In contrast, the error rates for SEM and BF were consistently elevated above 0.05, but declined with increasing sample size. SEM had the highest rates compared to any method. Please note again that these results combine the outcomes for all of the

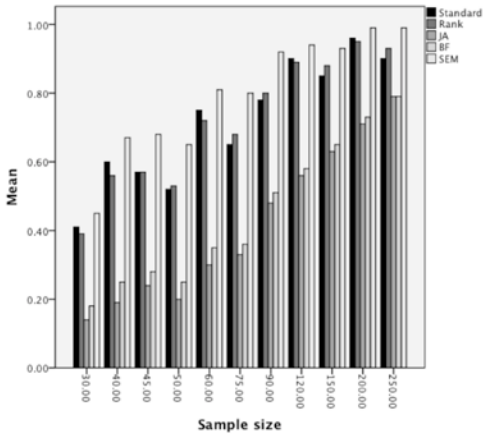
distributions, and that SEM maintained the nominal Type I error rate when the data were normally distributed, though it did not for the nonnormal data.

Distribution: Power

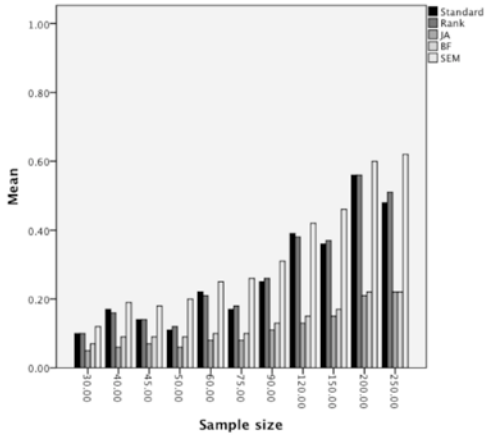
The factorial repeated ANOVA for the power of the MANOVA test statistics when the distributions were varied identified the interaction of method by correlation among dependent variables by distribution by number of variables ($p < 0.001$, $\omega^2 = 0.588$) and the interaction of method by sample size by effect size ($p < 0.001$, $\omega^2 = 0.694$) as the highest order significant terms with ω^2 greater than 0.10. All other significant lower order main effects and interactions were subsumed in these interactions and will not be discussed further.



Power for Large effect size



Power for Medium effect size



Power for Small effect size

Figure 3. Three panels of power rates for five MANOVA tests by effect size and sample size.

Figure 3 (three panels) displays power for the representative statistics used previously (standard, rank based, JA, BF and SEM) by effect size and sample size. When interpreting these results, it is important to keep in mind that the Type I error rates for SEM were inflated when the data were not normally distributed, and therefore higher power rates with SEM must be viewed with caution. The following discussion will focus on power for those statistics that maintained the nominal Type I error rate of 0.05. Across effect size and sample size values, the standard and rank based approaches maintained the highest power values of those methods that were able to maintain the nominal Type I error rate across distributions. In contrast, when the effect size was large, the BF and JA methods had lower power compared to the other approaches for the smallest sample size condition. Not until $N = 120$ did power approach 0.8 for these methods. When the simulated effect size was of medium magnitude, none of the methods that controlled Type I error had power rates approaching 0.8 until sample sizes were 90, and again the standard and rank approaches had higher power than JA or BF. In contrast, for the large effect condition the standard and rank statistics had markedly higher power rates across sample sizes, with values of 0.8 or greater for N of 60 or more. Finally, when the simulated effect size was small, the patterns were similar to those for larger effects, though none of the methods that controlled Type I error had power greater than 0.6 for any sample size, and the standard and rank based approaches had higher power than JA or BF.

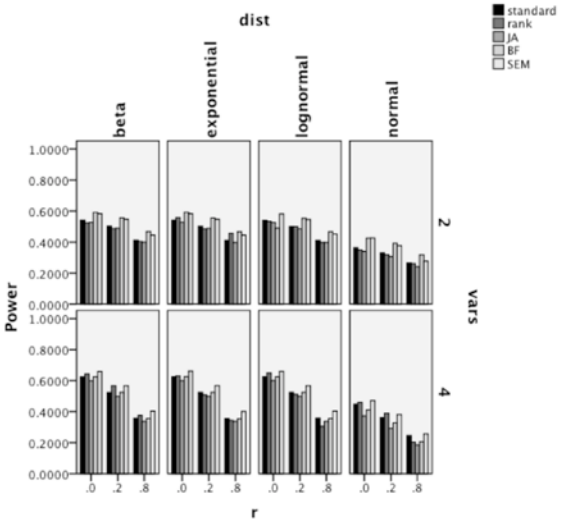


Figure 4. Power for five MANOVA tests by correlation (r) among dependent variables, distribution (dist) of dependent variables, and number of dependent variables (var)

Figure 4 includes the power rates for the significant 4-way interaction of test statistic by correlation among the dependent variables by distribution of the dependent variables by number of dependent variables. An examination of this figure reveals that across distributions and test statistics, power declined with increases in the correlation among the dependent variables. In terms of the test statistics that controlled Type I error, power for BF was generally the highest in the 2 variable case, and the standard, rank, and BF approaches displayed comparable power with 4 variables across distributions. With the normal distribution, SEM also had among the highest power values in the 4 variable condition, on par with standard and rank tests. And again, although SEM had the highest power values in most of the nonnormal conditions, it is not discussed in that context here due to the Type I error inflation it exhibited for the nonnormal distributions. JA consistently displayed among the lowest power results of the methods studied here. Power was consistently lower in the normal distribution condition compared to the other distributions studied here. Finally, note that power was below 0.80 across all conditions.

Discussion

The goals of this study were to provide a comprehensive review of the various test statistics available for MANOVA when standard assumptions are violated, and to conduct a large simulation study to compare the performance of the 16 identified (i.e., four families) test statistics across a variety of simulated conditions to evaluate Type I error and power. The results illustrate that Type I error and power do differ based on the selection of the test statistic for the MANOVA, dependent upon specific data conditions. This work is in accord with calls to make such comparisons. Raykov (2001), for example, encouraged comparison of the standard approach to testing the multivariate null hypothesis of no mean vector difference across groups as represented by *P* with an approach based upon SEM. This comparison was made, among several others, and extended this work to the 3 group case. Thus, this study does provide information on performance of these tests to aid the researcher in selecting the test statistic(s) that appears to work well given the data at hand, corresponding assumptions that are satisfied, and the variable framework (latent vs. observed) under which the analysis is conducted. Seven factors were manipulated resulting in 12,076 comparison conditions to gain a greater understanding of the relative performance of the standard approach for testing the multivariate hypotheses with respect to mean differences, along with a number of purportedly more robust options.

Major Points

Results revealed that when MANOVA assumptions are met, SEM and *P* are optimal in terms of Type I error and power rates. This result for *P* is consistent with prior research (e.g. Christensen & Rencher, 1997), though there is not a great deal of prior work examining many of the other alternative test statistics. Furthermore, both SEM and *P* maintained the nominal error rate in this condition, and SEM had the highest power rates. Even when data are not normally distributed, the *P* statistics maintain the nominal Type I error rate as do the Family 1 and Family 2 test statistics, thus partially supporting the first research hypothesis for this study. However, when the assumption of equal covariances is violated, but group sample sizes remains equal, the *P* statistic displays elevated Type I error rates whereas both SEM and Family 1 tests maintained the nominal rate. Moreover, the *P* statistic had severely inflated Type I error rates when the smaller group had the larger covariances. Again, both SEM and Family 1 test statistics were able to maintain the nominal error rate in this case. Family 3 performed similarly to the standard approach in terms of both Type I error rate and power in the case of three groups. However, for two groups, Family 3 had low power, making it of questionable utility under these conditions.

With regard to power under the unequal covariance matrix conditions, SEM, followed by the Family 1 tests, had the highest values compared to the other test statistics that were able to maintain the Type I error rate at or near the nominal 0.05 level. This positive performance for SEM is in keeping with Raykov's (2001) suggestion that this approach would be particularly useful when the group covariance matrices were not equivalent. When covariance matrices were unequal, the power rates of the standard statistic, or H, could not be fairly compared because their error rates were inflated, particularly when the smaller group was paired with the covariance matrix having the larger elemental values. H had inflated error rates across most conditions. In short, when the outcome variables followed the normal distribution, SEM was able to maintain the nominal Type I error rate, and yield higher rates of power than the other methods studied here. Furthermore, in accord with Raykov (2001), the SEM approach was optimal among all the methods when the group covariance matrices were not equal and the data were normally distributed. This result supports the expectation that by allowing the group covariance matrices to be independently estimated as in SEM, it is possible to produce accurate results even when the standard assumption of homogeneity of covariance matrices is not met.

Results for procedures using trimmed estimators were similar to those that used the usual least squares estimators, with slightly lower Type I error and power

rates compared to their non-trimmed counterparts. However, these differences were consistently very small, and generally did not offer a substantive advantage over the non-trimmed test statistics. Note that power for all methods was higher in the nonnormal conditions (no differences among these three) than for normal data. At the same time, there was no concomitant inflation of the Type I error rates for a number of the test statistics when non-normal data were present. The lack of influence of non-normality may be due to the adjustments that were examined. For instance, Hotelling's T^2 is conservative with skewed distributions or when outliers are in the tails of the distribution, especially when the design is unbalanced (Everitt, 1979; Zwick, 1986). It may be that under these conditions and with adjustments such as the use of the trimmed means, the other methods remain conservative as well. Lix & Keselman (2004) state that using Family 1 with the trimmed means can result in a test that is robust to the effects of both non-normality and covariance heterogeneity. When multivariate normality is violated, the performance of Hotelling's T^2, for example, can depend on the nature of the research design and the type of departure from normality present in the data. It appears this may be the case for the other tests as well. Furthermore, other findings have suggested it may be small sample sizes with non-normal data that result in liberal results or Type I error inflation (e.g., Fouladi & Yockey, 2002; Wilcox, 1995) with these studied test statistics. Such effects with various combinations of conditions appear to deserve continued investigation to assist in sorting out when one would and would not expect a degrading of statistical power or inflation of Type I error.

Given the relative success of the Family 1 tests, it may be beneficial to take a moment and reiterate how these differ from those of the other families. Recall from the earlier discussion of this issue that Family 1 are all based on $T^2_{unequal}$, which is an analog of the univariate *t-test* calculation when group variances differ. Thus, the variances are weighted by the inverse of the group size. For tests in Families 2 and 3, the weighting of group variances was based on more complex combinations of sample size or sample proportions. Thus, the use of a simple weighting of variances by the inverse sample size may be more effective than attempting to account for the proportion of total cases in the sample, for example. Furthermore, given the very similar performances of the statistics in Family 1 to one another, it seems that the alternative methods for calculating degrees of freedom that demarcate most of these may not be particularly meaningful in conditions similar to those simulated here.

The results of this study partially supported the hypothesis that the SEM approach would have lower Type I error and higher power for all but the smallest

sample sizes. When the underlying data were normally distributed, this method would seem to be a good choice for applied researchers. SEM consistently maintained Type I error control, and yielded the highest power values, regardless of whether the group covariance matrices were equal or not. The maximum likelihood based SEM approach is closely associated with the familiar Wilks' Lambda statistic, commonly used in MANOVA testing, when the data are normally distributed, with the exception that it can be used successfully when group variances are unequal. For nonnormal data distributions simulated here, SEM was not able to maintain the nominal error rate of 0.05.

Finally, results for the trimmed methods did not differ substantially from their non-trimmed counterparts, other than by exhibiting slightly lower rejection rates. The lack of higher power in the skewed case, which was hypothesized might occur, could be due to the fact that the data were not simulated to contain true outliers, given that this was not the focus. Thus, future research should include cases where outliers are present.

Practical Recommendations for Applied Researchers

The following guideline of bullet points summarizes results; these may prove to be helpful to researchers working with MANOVA in situations where the assumptions of normality and/or equality of covariance matrices are violated. These points are organized based upon the type of assumption violation and provide the researcher with suggested test statistics to use in each situation, based upon the results of this simulation study.

1) When data are normally distributed and the groups' covariance matrices are equal, SEM provides optimal power and Type I error control.

2) When the data are not normally distributed and the groups' covariance matrices are equal, the *P* statistic maintains the nominal Type I error rate and has optimal power, whereas SEM yields an inflated Type I error, and members of Family 1 do not.

3) When the groups' covariance matrices are not equal and data are normally distributed, the *P* statistic will exhibit an inflated Type I error rate, whereas SEM, and members of the Family 1 test statistics (except for H) will maintain the nominal error rate.

4) When the groups' covariance matrices are not equal and data are normally distributed, SEM will have the highest power rates, and the Family 1 test

statistics will have lower power to find group mean differences compared to the *P*.

5) Tests based on trimmed statistics demonstrated slightly lower Type I error rates and power than their non-trimmed analogs.

Study limitations and directions for future research

As with any simulation study, there are limitations to the current work. First, a limited number of covariance inequality conditions were considered in which values for one group were multiples of those for another. Future work should expand upon the current work by investigating other covariance structures. Second, for each distribution condition, the variables had the same distribution. In practice this may not be the case, and future research should simulate situations in which variables have different distributions from one another. Third, only three non-normal distributions were considered here. Further work could, for instance, examine heavy tailed symmetric distributions, such as the Cauchy. Finally, only positively correlated dependent variables were examined here. As was noted in the introduction, the presence of negative correlations among the responses can lead to increased power for MANOVA tests. Thus, future research could extend the current work by comparing the performance of several of these methods in the presence of negative dependent variable correlations.

Conclusion

There is little doubt that with sixteen options for test statistics for MANOVA, many researchers will be overwhelmed with the choice that must be made. Many applied researchers may even be completely unaware of the various choices that exist. Furthermore, many of the choices are not available as standard options in some commonly used statistical packages, which can hinder accurate as well as wide-spread use. The result of this relative lack of access is that valid hypothesis testing in multivariate means comparisons may not be obtained when assumptions underlying the hypotheses tests are not satisfied. However, the development of the SAS macro by Lix and Keselman (2004), as well as the increasing availability of easy to use and powerful software for SEM, make many of these alternatives more accessible than ever before. Therefore, the applied researcher is encouraged to carefully consider the selection of the test given data conditions and seek resources to assist in calculations of that statistic if need be. Developers of statistical software are also encouraged to continue to integrate the various

options of these test statistics even beyond MANOVA. Though there is likely to be a lag behind development of state-of-the-art methods and software to implement these methods, researchers are encouraged to continue to attempt the use of the most appropriate method or test given the data and research question at hand. It is anticipated that the review of test statistics and results of this study will assist in guiding applied researchers in selecting optimal methods for comparing multivariate group means.

References

Algina, J., Oshima, T. C. & Tang, K. L. (1991). Robustness of Yao's, James' and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. *Journal of Educational Statistics, 16,* 125-139.

Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. *Multivariate Behavioral Research, 29,* 141-163.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for Equality of Variances, *Journal of the American Statistical Association, 69,* 364–367.

Christensen, W. F., & Rencher, A. C., (1997). A comparison of Type I error rates and power levels for seven solutions to the multivariate Behrens–Fisher problem. *Communications in Statistics-Theory and Methods, 26,* 1251–1273.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115,* 465-474.

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. *Journal of Educational and Behavioral Statistics, 66,* 137-179.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Dumas, J. E., Prinz, R. J., Smith, P. E., & Laughlin, J. (1999). The EARLY ALLIANCE Prevention Trial: An Integrated Set of Interventions to Promote Competence and Reduce Risk for Conduct Disorder, Substance Abuse, and School Failure. *Clinical Child and Family Psychology Review, 2,* 37-53.

Erdfelder, E. (1981). Multivariate Rangvarianzanalyse: Ein non-parametrisches Analogonzurein- und mehrfaktoriellen MANOVA [Multivariate rank variance analysis: A nonparametric analogue for single and multivariate MANOVAs]. *TriererPsychologischeBerichte, 8.* Trier, German: Fachbereich 1-Psychologie derUniversitat Trier.

Everitt, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one and two sample T^2 tests. *Journal of the American Statistical Association, 74,* 48-51.

Finch, H. (2005). Comparison of the performance of the nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology, 1,* 27-38.

Finch, W. H., & French, B. F. (2008). Testing the null hypothesis of no group mean vector difference: A comparison of MANOVA and SEM. Paper presented at the Annual meeting of the Psychometric Society, Durham, NH, June.

Fouladi, R. T., &Yockey, R. D. (2002). Type I error control of two-group multivariate tests on means under conditions of heterogeneous correlation structure and varied multivariate distributions. *Communications in Statistics – Simulation and computation, 31,* 360-378.

Hair, J. F., Anderson, R. E., Tatham, R. L, & Black, W. C. (1987). *Multivariate data analysis with readings (3rd ed).* New York, NY, Macmillan. .

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T^2 procedure and the assumptionfor homogeneous covariance matrices. *Psychological Bulletin, 86,* 1255-1263.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2001). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling, 7,* 534-556

Harris, R. J., (2001). *A primer of multivariate statistics(3rd Ed).* Mahwah, NJ: Lawrence Erlbaum.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated non-normal distributions: Extending the Fleishman power method. *Psychometrika, 64,* 25-35.

Holloway, L. N., & Dunn, O. J. (1967). The robustness of Hotelling's T^2. *Journal of the American Statistical Association, 62,* 124-136.

Hopkins, J. W., & Clay, P. P. F. (1963). Some empirical distributions of bivariateT^2 and homoscedasticity criterion *M* under unequal variance and leptokurtosis. *Journal of the American Statistical Association, 58,* 1048-1053.

Huberty, C. L., & Morris, J. D., (1989). Multivariate analysis versus multiple univariate analysis, *Psychological Bulletin, 105,* 302-308.

Huberty, C. J, & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Hussein, A., & Carriere C. K. (2005) Group Sequential Procedures under Variance Heterogeneity. *Statistical Methods for Medical Research. 14*, 1-8.

Ittenbach, R. F., Chayer, D. E., Bruininks, R. H., Thurlow, M. L., & Beirne-Smith, M. (1993). Adjustment of young adults with mental retardation in community settings: comparison of parametric and nonparametric statistical techniques. *American Journal of Mental Retardation, 97,* 607-615.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratio of the population variances are unknown. *Biometrika, 41,* 19-43.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85-92.

Johnson, R.A. &Wichern, D.W. (2002). *Applied multivariate statistical analysis.* Upper Saddle River, NJ: Prentice Hall.

Katz, B. M.. & McSweeney, M. (1980). A multivariate Kruskal-Wallis test with post hoc procedures. *Multivariate Behavioral Research, 15,* 281-297.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika, 63,* 145-163.

Kim, S. & Olejnik, S. (2004). Bias and precision of multivariate effect size measures of association for a fixed-effect analysis of variance model. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April.

Kirk, R. E. (1995). *Experimental Design: Procedures for the behavioral sciences.* New York: Wadsworth Publishing.

Korin, B. P. (1972). Some comments on the homoscedasticity criterion *M* and the multivariate analysis of variance tests T^2, *W*, and *R*. *Biometrika, 59,* 215-216.

Krishnamoorthy, K., & Xia, Y. (2006). On selecting tests for equality of two normal mean vectors. *Multivariate Behavioral Research, 41,* 533-548.

Krull, V., Choi, S., Kirk, K., Prusick, L., & French, B. F. (2010). Lexical effects on spoken-word recognition in children with normal hearing. *Ear and Hearing, 31,* 102-114.

Lee, Y.-S. (1971). Asymptotic formulae for the distribution of a multivariate test statistic: Power comparisons of certain multivariate tests. *Biometrika,58,*647-651.

Lix, L. M., & Keselman, H. J. (2004). Multivariate tests of means in independent group designs: Effects of covariance heterogeneity and nonnormality. *Evaluation in the Health Professions, 27(1),* 45-69.

McCarroll, D., Crays, N., & Dunlap, W. P. (1992). Sequential ANOVAs and type I error rates, *Educational and Psychological Measurement, 52,* 387-393.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519-530.

Muthén, L. K., &Muthén, B. O. (1998–2004). *Mplus user's guide* (4thed.). Los Angeles: Muthén&Muthén.

Nel, D. G., & Van der Merwe, C.A., (1986). A solution to the Multivariate Behrens–Fisher problem. *Communications in Statistics-Theory and Methods, 15,* 3719–3735.

Olejnik, S. (2010). Multivariate analysis of variance. . In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods.* (pp. 328 - 328). NY: Routledge.

Olson, C.L. (1974). Comparative robustness of six test in multivariate analysis of variance. *Journal of the American Statistical Association, 69,* 894-908.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287-312.

Pillai, K. C. S., & Jayachandran, K. (1967) Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika, 54,*195-210.

Puri, M. L., & Sen, P. K. (1971). *Nonparametric methods in multivariate analysis.* Malabar, FL: Krieger Publishing Company.

Ramsey, P. H. (1982). Empirical power of procedures for comparing 2 groups on p variables. *Journal of Educational Statistics, 7,* 139-156.

Raykov, T. (2001). Testing multivariable covariance structure and means hypotheses via structural equation modeling. *Structural Equation Modeling, 8(2),* 224-256.

SAS Institute. (2004). SAS software version 9.1. Cary, NC: SAS Institute.

Satorra, A., & Bentler, P.M. (1994). *Corrections to test statistics and standard enors in covariance structure analysis.* In A. von Eye & C.C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp. 399-419). Thousand Oaks, CA: Sage.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika,40,* 87-104.

Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement, 58,* 861-881.

Stevens, J. (2001). *Applied Multivariate Statistics for the Behavioral Sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Warner**,** R. M. **(**2008**).** *Applied Statistics: From bivariate through multivariate techniques*. Thousand Oaks: Sage

Wilcox, R. R. (1995). Simulation results on solutions to the multivariate Behrens-Fisher problem via trimmed means. *The Statistician, 44,* 213-225.

Yanagihara, H., & Yuan, K-H. (2005). Three approximate solutions to the multivariate Behrens-fisher problem. *Communications in Statistics: Simulation and Computation, 34,* 957-988.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika, 52,* 139-147.

Yu, C., & Muthén, B. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Yuen, K. K. (1974). The two-stage sample trimmed *t* for unequal population variances. *Biometrika, 61,* 165-170.

Zwick, R. (1986). Rank and normal scores alternatives to Hotelling's T^2, *Multivariate Behavioral Research, 21,* 169-186

Appendix A

The below equations supplement the material in the text so the interested reader has the formulas at their disposal. The terms are defined below and correspond to terms which appear throughout the text. For addition information on the derivation of the statistics please see the cited sources in the text.

Family 1

1) **The multivariate analog of the univariate *t*-test equation for unequal variances:**

$$T^2_{unequal} = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} \left(\bar{Y}_1 - \bar{Y}_2\right)$$

2) ***F*$_{JN}$ involves the conversion of $T^2_{unequal}$ to an *F* value:**

$$F_{JN} = \frac{T^2_{unequal}}{c_2}$$

Where

$$c_2 = p + 2C - \frac{6C}{\left(p+1\right)}$$

p = Number of outcome variables

$$C = .5\sum_{j=1}^{2} \frac{1}{n_j} tr\left(A^{-1}A_j\right)^2 + tr^2\left(A^{-1}A_j\right)$$

$$A_j = \frac{S_j}{n_j}$$

$$A = A_1 + A_2$$

This F_{JN} value for this statistic is then compared with an F critical value p,v_J degrees of freedom with $v_J = p\ (p + 2) / 3C$.

3) The F_{NV} test statistic is a transformed version of $T^2_{unequal}$:

$$F_{NV} = \frac{v_N T^2_{unequal}}{pf_2}$$

Where

$$f_2 = \left(trA^2 + tr^2 A\right)\sum_{j=1}^{2}\frac{1}{n_j - 1}\left(trA_j^2 + tr^2 A_j\right)$$

$$v_N = f_2 - p + 1$$

F_{NV} is compared to a critical F value with p, v_N degrees of freedom.

4) Yao's F_Y is based on $T^2_{unequal}$:

$$F_Y = \frac{v_K T^2_{unequal}}{pf_1}$$

Where

$$f_1 = \sum_{j=1}^{2}\frac{1}{n_j - 1}\left(\frac{T^2_{unequal}}{b_j}\right)^2$$

$$b_j = \left(\bar{Y}_1 - \bar{Y}_2\right)' V^{-1} A_j V^{-1}\left(\bar{Y}_1 - \bar{Y}_2\right)$$

$$V = A_1 + r^2 A_2 + 2rA_2^{1/2} A_2 \left(A_2^{-1/2} A_1 A_2^{-1/2}\right) A_2^{1/2}$$

$$r = \left|A_1 A_2^{-1}\right|^{1/(2p)}$$

Family 2

5) The Brown and Forsythe (F_{BF}) test statistic:

$$F_{BF} = \frac{v_{BF2}}{pf_2} T_{BF}$$

Where

$$T_{BF} = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left[\left(1 - \frac{n_1}{N}\right)S_1 + \left(1 - \frac{n_2}{N}\right)S_2\right]^{-1}\left(\bar{Y}_1 - \bar{Y}_2\right)$$

$$v_{BF2} = f_2 - p + 1$$

$$f_2 = \frac{tr\left(G_1\right)^2 + tr^2\left(G_1\right)}{\frac{1}{n_{1-1}}\left[tr\left(w_1S_1\right)^2 + tr^2\left(w_1S_1\right)\right] + \frac{1}{n_{2-1}}\left[tr\left(w_2S_2\right)^2 + tr^2\left(w_2S_2\right)\right]}$$

$$w_j = 1 - \frac{n_j}{N}$$

$$G_1 = w_1S_1 + w_2S_2$$

$$v_{BF1} = \frac{tr\left(G_1\right)^2 + tr^2\left(G_1\right)}{tr\left(G_2\right)^2 + tr^2\left(G_2\right) + tr\left(\sqrt{w_1S_1}\right)^2 + tr\left(\sqrt{w_2S_2}\right)^2 + tr^2\left(\sqrt{w_1S_1}\right) + tr^2\left(\sqrt{w_2S_2}\right)}$$

$$G_2 = \frac{n_1}{N}S_1 + \frac{n_2}{N}S_2$$

6) The Kim (*FK*) test statistic:

$$F_K = \frac{v_k\left(\bar{Y}_1 - \bar{Y}_2\right)' V^{-1}\left(\bar{Y}_1 - \bar{Y}_2\right)}{c_1 m f_1}$$

Where

$$c_1 = \frac{\sum_{j=1}^{2} h_1^2}{\sum_{j=1}^{2} h_1}$$

$$h_1 = \frac{\left(d_1 + 1\right)}{\left(d_1^{1/2} + r\right)^2}$$

$$m = \frac{\left(\sum_{j=1}^{2} h_1\right)^2}{\sum_{j=1}^{2} h_1^2}$$

$$v_k = f_1 - p + 1$$

7) Winsorized variance:

$$S_{Wp}^2 = \frac{\sum_{i=1}^{n}\left(Z_i - \bar{Y}_{wp}\right)^2}{n-1}$$

Where

$\bar{Y}_{wp} =$ Winsorized mean of variable *p*

$Z_i = Y_{L+1}$ if $Y_i \le Y_L$

$Z_i = Y_{H-1}$ if $Y_i \geq Y_H$

Otherwise $Z_i = Y_i$

Y_L = Lower cut score corresponding to 20th percentile value.

Y_H = Upper cut score corresponding to 80th percentile value.

8) T^2 and $T^2_{unequal}$ **can be calculated using the trimmed means and Winsorized covariance matrices as:**

$$T_R^2 = \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)' \left[S_w \left(\frac{1}{h_1} + \frac{1}{h_2} \right) \right]^{-1} \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)$$

Where

$$S_w = \frac{(n_1 - 1)}{(h_1 - 1)} S_{w1} + \frac{(n_2 - 1)}{(h_2 - 1)} S_{w2}$$

$\bar{Y}_{Tj}$ = Trimmed mean for group *j*

h_j = Number of group *j* that is kept after trimming.

9) A version of Hotelling's T^2 that does not use the pooled covariance matrix:

$$T^2_{R\ unequal} = \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)' \left(\frac{(n_1 - 1)}{(h_1 - 1)h_1} S_{w1} + \frac{(n_2 - 1)}{(h_2 - 1)h_2} S_{w2} \right)^{-1} \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)$$

Family 3

10) Rank based nonparametric test

Convert Pillai's trace value using ranks into the chi-square statistic: $\chi^2 = (n-1)P$ where *P* is Pillai's trace and *n* is the total sample size. Compare the value with the χ^2 distribution with *k* (*p* – 1) degrees of freedom, where *k* is the number of groups for the independent variable and *p* is the number of response variables.

Family 4

11) Structural Equation Model based test

To test of the null hypothesis of group differences on the responses is the difference in the χ^2 fit statistics: $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}}$. Allowing the group means to differ results in a saturated CFA model so that the value of $\chi^2_{\text{Unconstrained}} = 0$.

The test of the null hypothesis of group differences across the vector of dependent variable means is equivalent to $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}} = \chi^2_{\text{Constrained}} - 0 = \chi^2_{\text{Constrained}}$.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 82-104.

ISSN 1538 − 9472

Constructing Confidence Intervals for Effect Sizes in ANOVA Designs

Li-Ting Chen
Indiana University
Bloomington, IN

Chao-Ying Joanne Peng
Indiana University
Bloomington, IN

A confidence interval for effect sizes provides a range of plausible population effect sizes (ES) that are consistent with data. This article defines an ES as a standardized linear contrast of means. The noncentral method, Bonett's method, and the bias-corrected and accelerated bootstrap method are illustrated for constructing the confidence interval for such an effect size. Results obtained from the three methods are discussed and interpretations of results are offered.

Keywords: Confidence interval, linear contrast, effect size, bootstrap, noncentral

Introduction

The importance of reporting effect sizes (ESs) and confidence intervals (CIs) has been strongly emphasized in the debate over null hypothesis significance testing as a methodology in social science research (Cohen, 1994; McCartney & Rosenthal, 2000; Nix & Barnette, 1998; Schmidt, 1996, although see Sawilowsky & Yoon, 2002, in this journal for a contrary view). Cumming (2012) characterized the shift from reliance on null hypothesis significance testing to the use of ESs, CIs, and meta-analyses as new statistics. Thompson (2002) stated, "An improved quantitative science would emphasize the use of confidence intervals (CIs), and especially CIs for effect sizes" (p.25), and constructing CIs for ESs facilitates meta-analytic thinking and interpretation. Thompson explained that reporting CIs allows future researchers to incorporate prior knowledge into the estimation of the same population ES. Furthermore, CI is directly related to the precision of ES estimates obtained from different studies. (See Knapp & Sawilowsky, 2001a, 2001b for a contrary view.)

Professional organizations such as the American Psychological Association (APA) and the American Educational Research Association (AERA) have both

Dr. Chen is a recent PhD graduate. Email her at: litchen@indiana.edu. Dr. Peng is a professor of inquiry methodology and adjunct professor of statistics in the Department of Counseling and Educational Psychology. Email her at: peng@indiana.edu.

stressed the importance of reporting CIs for ESs, particularly since 1999. According to the APA Task Force Report, "Interval estimates should be given for any effect sizes involving principal outcomes" (Wilkinson and the Task Force on Statistical Inference, 1999, p. 599). The fifth and sixth editions of the APA *Publication Manual* stress that "The inclusion of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results" (APA, 6th edition, 2010, p.34). In addition, the sixth edition of the APA *Publication Manual* emphasizes, "Whenever possible, provide confidence interval for each effect size reported to indicate the precision of estimation of the effect size" (APA, 6th edition, 2010, p.34). Likewise, the AERA's Standards for Reporting on Empirical Social Science Research suggest that, "For each of the statistical results that is crucial to the logic of the design and analysis, there should be included: … An indication of the uncertainty of that index of effect size …" (AERA, 2006, p. 37). According to the sixth edition of the APA Publication Manual, it is crucial to report confidence intervals because "confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy" (p. 34). For ways to report CIs, the same APA manual states, "As a rule, it is best to use a single confidence level, specified on an a priori basis (e.g., a 95% or 99% confidence interval), throughout the manuscript. Wherever possible, base discussion and interpretation of results on point and interval estimates" (p. 34).

Despite these efforts, the reporting rate of CIs for ESs in empirical studies is still low (Odgaard & Fowler, 2010; Peng, Chen, Chiang, & Chiang, 2013). This phenomenon may be due to a lack of understanding of the statistical properties of CIs for ESs, or a lack of suitable algorithms for the construction of CIs implemented in commercial statistic software (e.g., SPSS, SAS). Thus, this article aims to present three methods and algorithms for constructing the CI for a standardized linear contrast of means in a one-way fixed-effects univariate ANOVA design. This article defines a standardized linear contrast of means as a measure of ES for fixed-effects ANOVA designs. And the three methods are: the noncentral method, Bonett's method, and the bias-corrected and accelerated bootstrap method.

To facilitate the understanding of standardized linear contrasts of means and to illustrate the three methods, a sleep deprivation example from Kirk (1995) is used. This example serves as a template for discussing the construction of CI for a standardized linear contrast of means using the three methods.

A sleep deprivation example

This example examines the effects of sleep deprivation on hand-steadiness. According to Kirk (1995): Assume an interest in the effects of sleep deprivation, treatment *A*, on hand-steadiness. The four levels of sleep deprivation of interest are 12, 18, 24, and 30 hours, which are denoted by a_1, a_2, a_3, and a_4, respectively. An experiment is conducted in which 32 subjects are randomly assigned to the four levels of sleep deprivation, with the restriction that eight subjects are assigned to each level. The dependent variable is the number of time during a 2-minute interval that a stylus makes contact with the side of a ½-inch hole (p. 166). The independent variable is hours of sleep deprivation and the dependent variable is the number of times that a stylus held by a participant makes contact with the side of a ½-inch hole. The higher the number, the worse the performance, presumably affected by the deprivation of sleep. Data gathered from this study are shown in Table 1.

Table 1. The number of times that a stylus held by a participant makes contact with a ½-inch hole during a 2-minute interval from the sleep deprivation sample.

Hours of Sleep Deprivation	**12 hours**	**18 hours**	**24 hours**	**30 hours**
Treatment Level	a_1	a_2	a_3	a_4
	4	4	5	3
	6	5	6	5
	3	4	5	6
	3	3	4	5
	1	2	3	6
	3	3	4	7
	2	4	3	8
	2	3	4	10
Group Sizes (n_j)	8	8	8	8
Group Means ($\overline{Y}_j$)	3	3.5	4.25	6.25
Standard deviation ($\hat{\sigma}_j$)	1.51	0.93	1.04	2.12

Consider the hypothesis that a human's fine motor skill decreases dramatically after being deprived of sleep for 24 hours or longer. Thus, interest lies in the contrast between the average performance of participants after 24 and

30 hours of sleep deprivation versus the average performance of participants after 12 and 18 hours. The linear contrast of means (ψ) is written as

$$\psi = 0.5\times\left(\mu_{24\text{ hours}} + \mu_{30\text{ hours}}\right) - 0.5\times\left(\mu_{12\text{ hours}} + \mu_{18\text{ hours}}\right) = \sum_{j=1}^{k} c_j \mu_j, \quad (1)$$

where μ_j is the population mean for the jth group, k is the number of independent groups (= 4 for the sleep deprivation example), and c_j is the coefficient or weight assigned to the jth group (= 0.5, 0.5,− 0.5, and −0.5 for 24 hours, 30 hours, 12 hours, and 18 hours of sleep deprivation, respectively). Equation 1 and all subsequent equations are written specifically to suit the sleep deprivation example first, followed by a general formulation (in blue).

The value obtained from Equation 1 based on sampled data is an estimate of the corresponding population ES in original units. If a researcher wishes to standardize this ES, he/she needs to divide ψ with a standardizer. Such a standardizer is usually the population standard deviation, assumed to be equal and expressed as σ. For the specific ψ defined in Equation 1, its standardized form (δ) is written as follows:

$$\delta = \frac{\psi}{\sigma} = \frac{0.5\times\left(\mu_{24\text{ hours}} + \mu_{30\text{ hours}}\right) - 0.5\times\left(\mu_{12\text{ hours}} + \mu_{18\text{ hours}}\right)}{\sigma} = \frac{\sum_{j=1}^{k} c_j \mu_j}{\sigma}. \quad (2)$$

Reporting a standardized linear contrast of means is more informative than reporting a linear contrast of means in original units, when (1) the original unit of the dependent variable is not familiar to readers, or (2) a researcher intends to compare ESs obtained from studies that employ different dependent variables.

The following three sections introduce three methods for constructing CIs for standardized linear contrasts of means as ESs. The three methods are the noncentral method, Bonett's method, and the BCa (or the bias-corrected and accelerated bootstrap) method. After obtaining CIs results are compared and proper interpretations of CIs in this context are discussed.

Methods

Noncentral Method

Within the null hypothesis significance testing framework, a linear contrast ψ defined in Equation 1 is tested with a t-statistic defined as:

$$t = \frac{0.5\times\left(\overline{Y}_{24\text{ hours}} + \overline{Y}_{30\text{ hours}}\right) - 0.5\times\left(\overline{Y}_{12\text{ hours}} + \overline{Y}_{18\text{ hours}}\right)}{\hat{\sigma}\times\sqrt{\frac{(0.5)^2 + (0.5)^2 + (-0.5)^2 + (-0.5)^2}{8}}} = \frac{\sum_{j=1}^{k} c_j \overline{Y}_j}{\hat{\sigma}\times\sqrt{\sum_{j=1}^{k}\frac{c_j^2}{n_j}}}, \qquad (3)$$

where $\overline{Y}_j$ is the sample mean for the *j*th group (= 4.25, 6.25, 3.00, and 3.50 for 24, 30, 12, and 18 hours of sleep deprivation, respectively), $\hat{\sigma}$ is the pooled standard deviation that is used to estimate the equal population standard deviation $(=\sqrt{\left(1.51^2 + 0.93^2 + 1.04^2 + 2.12^2\right)/4} = 1.48)$, and n_j is the sample size for the *j*th group ($=\sqrt{8}$ for each of the four groups in sleep deprivation example).

Under the null hypothesis of a 0 linear contrast of means, the *t* statistic is distributed as a symmetric central *t* distribution with a mean of 0. When the null hypothesis is false (meaning the population linear contrast of means does not equal 0), the *t* statistic follows a noncentral *t* distribution that is centered approximately at the noncentrality parameter λ, when the degree of freedom is large (see Cumming & Finch, 2001). The noncentral *t* distribution has two parameters: the degrees of freedom (or *df* = the number of participants − the number of independent groups) and λ. When λ is zero, the noncentral *t* distribution is the central *t* distribution, or simply the *t* distribution.

One way to construct the CI for δ defined in Equation 2, is to use the noncentral *t* distribution. The noncentrality parameter λ of the noncentral *t* distribution is related to δ as follows,

$$\delta = \lambda\times\sqrt{\frac{0.5^2}{8} + \frac{0.5^2}{8} + \frac{(-0.5)^2}{8} + \frac{(-0.5)^2}{8}} = \lambda\times\sqrt{\sum_{j=1}^{k}\frac{c_j^2}{n_j}}. \qquad (4)$$

And λ is defined as follows,

$$\lambda = \frac{0.5\times\left(\mu_{24\text{ hours}} + \mu_{30\text{ hours}}\right) - 0.5\times\left(\mu_{12\text{ hours}} + \mu_{18\text{ hours}}\right)}{\sigma\times\sqrt{\frac{0.5^2}{8} + \frac{0.5^2}{8} + \frac{(-0.5)^2}{8} + \frac{(-0.5)^2}{8}}} = \frac{\delta}{\sqrt{\frac{0.5^2}{8} + \frac{0.5^2}{8} + \frac{(-0.5)^2}{8} + \frac{(-0.5)^2}{8}}} = \frac{\sum_{j=1}^{k} c_j \mu_j}{\sigma\times\sqrt{\sum_{j=1}^{k}\frac{c_j^2}{n_j}}}. \qquad (5)$$

Steiger and Fouladi (1997) illustrated how to derive λ from the observed t statistic obtained from a sample. From λ, using Equation 4, δ can be derived. To construct a 95% confidence interval for δ, first compute the lower and the upper limits of λ from the observed t statistic. The lower limit for λ is the noncentrality parameter of the noncentral t distribution in which the observed t statistic is at the 97.5th percentile. The upper limit for λ is the noncentrality parameter of the noncentral t distribution in which the observed t statistic is at the 2.5th percentile. From the two limits of λ, the limits for δ can be derived.

The use of noncentral distributions in constructing the CI for ESs involves sequence of iterations. In recent years, the computational difficulty for the noncentral t distribution has been overcome by algorithmic improvement. For example, the lower and upper limits of λ can be obtained in SAS® with the following syntax:

lamda_lower=TNONCT (t_observed, *df* , .975);

and

lamda_upper=TNONCT (t_observed, *df* , .025);

The *df* for the current example is 32 − 4 = 28. Once the lower limit and the upper limit of λ are obtained from SAS®, the lower limit and the upper limit of δ can be computed from the following according to Equation 4:

$$\delta_{\text{lower}} = \lambda_{\text{lower}} \times \sqrt{\frac{0.5^2}{8} + \frac{0.5^2}{8} + \frac{(-0.5)^2}{8} + \frac{(-0.5)^2}{8}} = \lambda_{\text{lower}} \times \sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}, \text{ and} \quad (6)$$

$$\delta_{\text{upper}} = \lambda_{\text{upper}} \times \sqrt{\frac{0.5^2}{8} + \frac{0.5^2}{8} + \frac{(-0.5)^2}{8} + \frac{(-0.5)^2}{8}} = \lambda_{\text{upper}} \times \sqrt{\sum_{j=1}^{k} \frac{c_j^2}{n_j}}. \quad (7)$$

Applying the noncentral method for constructing the CI for a standardized linear contrast of means is discussed in the literature (Cumming & Finch, 2001; Kline, 2004; Steiger, 2004). Liu (2010) illustrated the geometric meaning of the noncentrality parameter for a linear contrast in a Euclidian space. Kelley and Rausch (2006) and Lai and Kelley (2012) considered the sample size required to achieve the desired accuracy in CI estimations. Readers should note that there are

two statistical assumptions associated with the noncentral confidence intervals for δ. These two assumptions are (1) normality for each population distribution and (2) equal variances for all population distributions.

The SAS® macro "cinoncentral" (See Appendix A) yields the noncentral CI for a standardized linear contrast of means (δ). To execute this SAS® macro, readers first create a SAS data set in the DATA step of SAS®, or import the data into SAS®. This step is followed by the specification of a level of confidence, such as .95, and a coefficient for each group.

Bonett's Method

Bonett (2008) proposed a more general definition of the standardized linear contrast of means in order to deal with unequal variances across populations:

$$\delta_{\text{Bonett}} = \frac{0.5\times\left(\mu_{24\text{ hours}} + \mu_{30\text{ hours}}\right) - 0.5\times\left(\mu_{12\text{ hours}} + \mu_{18\text{ hours}}\right)}{\sqrt{\frac{\sigma^2_{24\text{ hours}} + \sigma^2_{30\text{ hours}} + \sigma^2_{12\text{ hours}} + \sigma^2_{18\text{ hours}}}{4}}} = \frac{\sum_{j=1}^{k} c_j\mu_j}{\sqrt{\frac{\sum_{j=1}^{k}\sigma_j^2}{k}}} = \frac{\psi}{\sqrt{\frac{\sum_{j=1}^{k}\sigma_j^2}{k}}}. \quad (8)$$

It is worth noting that, when population variances are equal (i.e., $\sigma^2_{24\text{ hours}} = \sigma^2_{30\text{ hours}} = \sigma^2_{12\text{ hours}} = \sigma^2_{18\text{ hours}} = \sigma^2$),

$$\delta_{\text{Bonett}} = \frac{0.5\times\left(\mu_{24\text{ hours}} + \mu_{30\text{ hours}}\right) - 0.5\times\left(\mu_{12\text{ hours}} + \mu_{18\text{ hours}}\right)}{\sqrt{\frac{4\sigma^2}{4}}} = \frac{0.5\times\left(\mu_{24\text{ hours}} + \mu_{30\text{ hours}}\right) - 0.5\times\left(\mu_{12\text{ hours}} + \mu_{18\text{ hours}}\right)}{\sigma} = \frac{\sum_{j=1}^{k} c_j\mu_j}{\sigma} = \delta. \quad (9)$$

In other words, δ is a special case of δ_{Bonett} when population variances are all equal. Based on a large sample approximation, Bonett derived the CI for δ_{Bonett} as follows:

$$\hat{\sigma}_{\text{Bonett}} \pm z_{\text{critical}}\left[\text{var}\left(\hat{\delta}_{\text{Bonett}}\right)\right]^{1/2}, \quad (10)$$

where z_{critical} is the critical value from the standard normal distribution, $\hat{\delta}_{\text{Bonett}}$ is the sample estimate for the corresponding population δ_{Bonett}, and $\text{var}\left(\hat{\delta}_{\text{Bonett}}\right)$ is

the sample variance of $\hat{\delta}_{\text{Bonett}}$. The sample estimate for Bonett's standardized linear contrast of means, i.e., $\hat{\delta}_{\text{Bonett}}$, is obtained from the following equation:

$$\hat{\delta}_{\text{Bonett}} = \frac{0.5\times\left(\overline{Y}_{24\text{ hours}} + \overline{Y}_{30\text{ hours}}\right) - 0.5\times\left(\overline{Y}_{12\text{ hours}} + \overline{Y}_{18\text{ hours}}\right)}{\sqrt{\frac{\hat{\sigma}^2_{24\text{ hours}} + \hat{\sigma}^2_{30\text{ hours}} + \hat{\sigma}^2_{12\text{ hours}} + \hat{\sigma}^2_{18\text{ hours}}}{4}}} = \frac{\sum_{j=1}^{k} c_j \overline{Y}_j}{\sqrt{\frac{\sum_{j=1}^{k} \hat{\sigma}^2_j}{k}}} = \frac{\sum_{j=1}^{k} c_j \overline{Y}_j}{\hat{\sigma}_{\text{Bonett}}}. \quad (11)$$

It is worth noting that when sample sizes are all equal, $\hat{\sigma}_{\text{Bonett}} = \hat{\sigma}$. And $\text{var}\left(\hat{\delta}_{\text{Bonett}}\right)$ is obtained from the following equation:

$$\text{var}\left(\hat{\delta}_{\text{Bonett}}\right) = \left(\frac{\hat{\delta}^2_{\text{Bonett}}}{4^2\hat{\sigma}^4_{\text{Bonett}}}\right)\times\left[\frac{\hat{\sigma}^4_{24\text{ hours}}}{2\times(8-1)} + \frac{\hat{\sigma}^4_{30\text{ hours}}}{2\times(8-1)} + \frac{\hat{\sigma}^4_{12\text{ hours}}}{2\times(8-1)} + \frac{\hat{\sigma}^4_{18\text{ hours}}}{2\times(8-1)}\right]$$
$$+\frac{\left[\frac{(0.5)^2\times\hat{\sigma}^2_{24\text{ hours}}}{(8-1)} + \frac{(0.5)^2\times\hat{\sigma}^2_{30\text{ hours}}}{(8-1)} + \frac{(-0.5)^2\times\hat{\sigma}^2_{12\text{ hours}}}{(8-1)} + \frac{(-0.5)^2\times\hat{\sigma}^2_{18\text{ hours}}}{(8-1)}\right]}{\hat{\sigma}^2_{\text{Bonett}}} \quad (12)$$
$$= \left(\frac{\hat{\delta}^2_{\text{Bonett}}}{k^2\hat{\sigma}^4_{\text{Bonett}}}\right)\sum_{j=1}^{k}\frac{\hat{\sigma}^4_j}{2\times df_j} + \frac{\sum_{j=1}^{k}\frac{c_j^2\hat{\sigma}^2_j}{df_j}}{\hat{\sigma}^2_{\text{Bonett}}},$$

where df_j = the number of participants in the jth group minus 1.

Bonett's method assumes normality, but not equal variances for the population distributions. When population variances are equal, δ becomes a special case of δ_{Bonett}. The SAS® macro "cibonett" (See Appendix B) yields Bonett's CI for a standardized linear contrast of means (δ_{Bonett}). To execute this SAS® macro, readers first create a SAS data set in the DATA step of SAS®, or import the data into SAS®. This step is followed by the specification of a level of confidence, such as .95, and a coefficient for each group.

The BCa Bootstrap Method

The bootstrap method is a resampling technique that constructs an empirical distribution of estimates from data already collected. Thus, the bootstrap method does not require assumptions of either normality or equal variances. Nor does it rely on a theoretical sampling distribution, such as *t* or normal, to derive the lower or the upper limits of a confidence interval.

Several methods of constructing CIs based on bootstrapping have been developed. These include the symmetric percentile bootstrap method, the bias-corrected and accelerated (BCa) bootstrap method, and the approximate bootstrap confidence (ABC) interval method. The BCa bootstrap method introduced here corrects the bias in the symmetric percentile bootstrap method. To provide the general idea of bootstrapping technique, the symmetric percentile bootstrap method is presented first, followed by the BCa bootstrap method.

The symmetric percentile bootstrap method constructs the CI by finding the $\alpha/2\times B$th and $[1-(\alpha/2)]\times B$th ranked values of the empirical distribution of the sample estimates. Here, α is the Type I error rate, such as .05; B is the number of bootstrap samples, such as 1,000. A bootstrap sample is a random sample of size n, drawn with replacement from the observed n scores. After a large number of bootstrap samples (e.g., 1,000) are formed, an empirical bootstrap distribution of the estimated effect sizes is constructed. From the empirical bootstrap distribution, the lower and upper confidence limits are derived. If $\alpha = 0.05$ and $B = 1{,}000$, the lower limit of a 95% bootstrap confidence interval is the $0.05/2\times 1{,}000^{\text{th}}$ ranked value of the empirical bootstrap distribution and the upper limit is the $[1-(0.05/2)]\times 1{,}000^{\text{th}}$ ranked value.

Readers can apply the bootstrap technique to construct the CI for either δ (Equation 2) or δ_{Bonett} (Equation 8). A step-by-step instruction for obtaining the CI for δ_{Bonett} using the symmetric percentile bootstrap method is presented. Readers can use this instruction to construct the CI for δ as well. Assuming that a 95% CI is to be constructed for δ_{Bonett}, based on the sample estimator $\hat{\delta}_{\text{Bonett}}$. Three steps are completed B times (e.g., 1,000) in order to yield B bootstrap estimates, $\hat{\delta}^{*}_{\text{Bonett}}$. The first step is to randomly sample eight scores with replacement from each of the four groups, where eight is the number of scores in each group. The second step is to compute the mean and the standard deviation of these eight scores from each group. The third step is to compute $\hat{\delta}^{*}_{\text{Bonett}}$ by plugging the means and the standard deviations obtained from the second step into Equation 11. After obtaining the B (e.g., 1,000) bootstrap $\hat{\delta}^{*}_{\text{Bonett}}$, one lists them in an ascending order.

The .025 × *B*th (=25th) and the .975 × *B*th (=975th) ranked $\hat{\delta}^{*}_{\text{Bonett}}$ are, respectively, the 95% lower and upper confidence limits. When estimates are tied, an average rank is assigned to the tied estimates.

The BCa bootstrap method is an improvement over the symmetric percentile bootstrap method. Specifically, it constructs the CI for the standardized linear contrast of means (δ or δ_{Bonett}) using $\text{CI}_{\text{lower}} \times B$th and $\text{CI}_{\text{upper}} \times B$th ranked values of bootstrapped estimates. The values of CI_{lower} and CI_{upper} depend on acceleration and bias-correction numbers, $\hat{a}$ and $\hat{z}_0$, respectively. According to Efron and Tibshirani (1993), $\hat{a}$ refers to the rate of change in the standard error of the estimated parameter (i.e., $\hat{\delta}_{\text{Bonett}}$) with respect to the true population value (i.e., δ_{Bonett}). The bias-correction number, $\hat{z}_0$, is interpreted as the median bias of the sample bootstrapped estimates. When exactly 50% of bootstrapped estimates are less than or equal to the observed estimate, $\hat{z}_0 = 0$. Using the notations described above for the symmetric percentile bootstrap method,

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#\left(\hat{\delta}^{*}_{\text{Bonett}} < \hat{\delta}_{\text{Bonett}}\right)}{B\left(=1{,}000\right)}\right], \tag{13}$$

where Φ^{-1} is the inverse of the standard normal cumulative function and $\#\left(\hat{\delta}^{*}_{\text{Bonett}} < \hat{\delta}_{\text{Bonett}}\right)$ is the frequency of those bootstrap estimates (i.e., $\hat{\delta}^{*}_{\text{Bonett}}$) that are less than the observed estimate (i.e., $\hat{\delta}_{\text{Bonett}}$). The acceleration $\hat{a}$ is obtained using the jackknife method that takes the form:

$$\hat{a} = \frac{\sum_{i=1}^{32}\left(\overline{\hat{\delta}}_{\text{Bonett}} - \hat{\delta}_{\text{Bonett}-i}\right)^{3}}{6\left\{\sum_{i=1}^{32}\left(\overline{\hat{\delta}}_{\text{Bonett}} - \hat{\delta}_{\text{Bonett}-i}\right)^{2}\right\}^{\frac{3}{2}}}, \tag{14}$$

where $\hat{\delta}_{\text{Bonett}-i}$ is the value of $\hat{\delta}_{\text{Bonett}}$ with the *i*th score removed from the entire data (*i* = 1,…,32 in the sleep deprivation example), and $\overline{\hat{\delta}}_{\text{Bonett}}$ is the average of all possible $\hat{\delta}_{\text{Bonett}-i}$. The CI_{lower} and CI_{upper} for a 95% confidence interval are given by

$$\mathrm{CI}_{\text{lower}} = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{\left(\frac{\alpha}{2}=.025\right)}}{1-\hat{a}\left(\hat{z}_0 + z^{\left(\frac{\alpha}{2}=.025\right)}\right)}\right], \text{ and} \tag{15}$$

$$\mathrm{CI}_{\text{upper}} = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{\left(1-\frac{\alpha}{2}=.975\right)}}{1-\hat{a}\left(\hat{z}_0 + z^{\left(1-\frac{\alpha}{2}=.975\right)}\right)}\right]. \tag{16}$$

Here, $z^{\left(\frac{\alpha}{2}\right)}$ is the $(100\times\frac{\alpha}{2})$th percentile of a standard normal distribution. For a 95% confidence interval, $z^{\left(\frac{\alpha}{2}=.025\right)} = -1.96$ and $z^{\left(1-\frac{\alpha}{2}=.975\right)} = 1.96$. The BCa confidence intervals yield the same results as the symmetric percentile bootstrap confidence intervals, when $\hat{z}_0$ and $\hat{a}$ both equal 0. In other words, $\mathrm{CI}_{\text{lower}} = \Phi(-1.96) = .025$ and $\mathrm{CI}_{\text{upper}} = \Phi(1.96) = .975$, when $\hat{z}_0$ and $\hat{a}$ both equal 0. The BCa method is superior to the symmetric percentile bootstrap method because it leads to better approximations to the lower and upper limits. However, Efron and Tibshirani (1993) stated, "their [the BCa] coverage accuracy can still be erratic for small sample sizes" (p.178). Chen's dissertation (2013) uncovered that the coverage probability produced by the BCa method was satisfactory when each group size was 30. Kelley's (2005) simulation found that BCa method's coverage probability was poor when each group size was eight.

The process for constructing the BCa CI for δ_{Bonett} may appear complex to some readers. However, a SAS® macro "cibca" (See Appendix C) based on the SAS program written by Barker (2005) is provided here to assist researchers in constructing BCa CIs for δ_{Bonett}, of which δ is a special case. To execute this SAS® macro, readers first create a SAS data set in the DATA step of SAS®, or import the data into SAS®. This step is followed by the specification of a number of bootstrap estimates (e.g., 1,000), a coefficient for each group, and a level of confidence, such as .95.

Results

Given the standardized linear contrast of means (δ in Equation 2) from the sleep deprivation example, the 95% noncentral CI was computed to be [0.56758, 2.12318]. Likewise, given the standardized linear contrast of means (δ_{Bonett} in Equation 8), the 95% Bonett's CI was computed to be [0.50045, 2.20958] and the BCa bootstrap CI to be [0.51967, 2.13941]. Thus, the CIs constructed by the three methods are slightly different from each other. The noncentral CI is the narrowest (= 2.12318 − 0.56758 = 1.55560) or most precise, followed by the BCa bootstrap CI (= 2.13941− 0.51967 = 1.61973), and the Bonett's CI (= 2.20958 −0.50045 = 1.70913). These results are consistent with findings obtained by Chen (2013) in a thorough investigation of these three methods under a variety of conditions.

In actuality, it is not necessary to compute more than one CI for a standardized ES. It is however necessary for researchers to be informed of the optimal method for a particular research context. For the purpose of demonstration, the correct interpretation of the noncentral CI for the contrast of interest (i.e., δ in Equation 2) is described.

How to interpret confidence intervals for a standardized linear contrast of means?

The 95% noncentral CI ranges from 0.56758 (or 0.57) to 2.12318 (or 2.12). Derived from the data presented in Table 1, all values contained in this interval cannot be rejected with a Type I error rate of 5%, if they are placed in a null hypothesis. Furthermore, all values in this interval are greater than 0. Thus, a null hypothesis of 0 standardized mean difference should be rejected at an α level of .05. Based on the data and the noncentral CI, readers can conclude that the difference in the number of times that a stylus touched the sides of a ½ hole between people deprived of sleep for 24 hours or longer and people deprived of sleep less than 24 hours can be as large as twice of the standard deviation of the data, or as small as a half of the standard deviation.

Discussion

A measure of an ES gives a point estimate of a treatment effect, whereas a CI of such an ES provides the precision of the estimation. Although both the APA and the AERA have encouraged researchers to report CIs for ESs, Odgaard and Fowler's (2010) study found that the reporting rate of CIs for ESs was only 40% in the *Journal of Consulting and Clinical Psychology*—the first APA journal that

required the reporting of CIs for ESs for primary results. A similar finding is reported by Peng et al. (2013) across a variety of psychology and education journals. The reporting of CI can be encouraged through accessible and reliable computing algorithms.

This article (1) illustrates the need to report CIs for ESs, (2) addresses the importance of reporting the CIs for ESs, (3) introduces, demonstrates, and compares three methods (the noncentral method, Bonett's method, and the BCa bootstrap method) for constructing the CI for a standardized linear contrast of means (a measure of the ES), and (4) provides SAS programming codes for these methods. The readers should note that the SAS programming codes provided in Appendices A – C are applicable for unequal sample sizes as well. It is hoped that this paper facilitates researchers' understanding of these three methods and enables them to report the CIs for ESs, defined as standardized linear contrasts of means in fixed-effects ANOVA designs.

Acknowledgements

This research was supported in part by the Maris M. Proffitt and Mary Higgins Proffitt Endowment Grant of Indiana University, awarded to the second author while the first author worked on the project as a research assistant.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40. doi: 10.3102/0013189X035006033

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Barker, N. (2005). A practical introduction to the bootstrap using the SAS system. *Proceedings of SAS conference: Phuse*. Retrieved from http://www.lexjansen.com/phuse/2005/pk/pk02.pdf

Bonett, D.G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13*, 99-109. doi: 10.1037/1082-989X.13.2.99

Chen, L.-T. (2013). *Effect size measures and their interval estimations: The multi-independent group case.* (Doctoral dissertation). Indiana University, Bloomington.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003. doi: 10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge/Taylor & Francis Group.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574. doi: 10.1177/0013164401614002

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*, 363-385. doi: 10.1037/1082-989x.11.4.363

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65, 51-69. doi: 10.1177/0013164404264850

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kline, R. B. (2004). *Beyond significance testing*. Washing, DC: American Psychological Association.

Knapp, T. R., & Sawilowsky, S. S. (2001a). Strong arguments: Rejoinder to Thompson. *Journal of Experimental Education, 70*, 94-95.

Knapp, T. R., & Sawilowsky, S. S. (2001b). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education, 70*, 65-79.

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *British Journal of Mathematical & Statistical Psychology, 65*, 350-370. doi: 10.1111/j.2044-8317.2011.02029.x

Liu, X. S. (2010). A note on noncentrality parameters for contrast tests in a one-way analysis of variance. *The Journal of Experimental Education, 78*, 53-59. doi: 10.1080/00220970903224669

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development, 71*, 173-180. doi: 10.1111/1467-8624.00131

Nix, T. W., & Barnette, J. J. (1998). The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing. *Research in the Schools, 5*, 3-14.

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology, 78*, 287-297. doi: 10.1037/a0019294

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25*, 157-209. doi: 10.1007/s10648-013-9218-2

Sawilowsky, S. S., & Yoon, J. S. (2002). The trouble with trivials. *Journal of Modern Applied Statistical Methods, 1*(1), 143-144.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129. doi: 10.1037/1082-989X.1.2.115

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*, 164-182. doi: 10.1037/1082-989X.9.2.164

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Hillsdale, NJ: Erlbaum.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32. doi: 10.3102/0013189X031003025

Wilkinson, L., and the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals - Guidelines and explanations. *American Psychologist, 54*, 594-604. doi: 10.1037//0003-066X.54.8.594

Appendix A

SAS® Macro “cinoncentral”

```
*************************************************************************;
*   format for data set: first variable = group, second variable = scores  ;
*   data = data set name                                                   ;
*   nominal = nominal confidence level                                     ;
*   contrast = a coefficient for each group, e.g., {.5 .5 -.5 -.5}         ;
*************************************************************************;
%MACRO cinoncentral(data,nominal,contrast);

PROC IML;
USE &data;
READ ALL INTO datain;
nn=NROW(datain);
groups = UNIQUE(datain[,1]);        *get group information of the data;
ngroups = NCOL(groups);             *get number of groups of the data;
CALL SYMPUTX("n_groups",ngroups);   *set the macro variable n_groups to be the
                                     number of groups;
%DO i = 1 %TO &n_groups;            *loop for groups;
group&i = datain[LOC(datain[,1]=&i),2]; /*obtain all the scores for each group*/
mu&i=mean(group&i);
sum&i = sum(group&i);
v_sum&i =(sum&i)**2;
n&i = nrow(group&i);
v_sum_n&i=v_sum&i/ n&i;
%END;
mu=mu1;
v_sum=v_sum1;
n=n1;
v_sum_n=v_sum1/n1;
%DO i=2 %TO &n_groups;
mu=mu//mu&i;
v_sum = v_sum//v_sum&i;
n = n//n&i;
v_sum_n=v_sum_n//v_sum_n&i;
contrast = t(&contrast);
%END;
df=n-1;
numerator=(contrast)`*mu;
mse1=(datain[,2])`*(datain[,2]); *squared values of all scores;
mse2=sum(v_sum_n);
mse=(mse1-mse2)/(nn-ngroups);
contrast_square=(contrast)##2;
n_1=1/n;
nu=(contrast_square)`*(n_1);
t=numerator/(SQRT(mse*nu));
lamda_lower = TNONCT(t,nn-ngroups,1-(1-&nominal)/2);  /*compute the lower
                                                        noncentrality*/
lamda_upper = TNONCT(t,nn-ngroups,(1-&nominal)/2);  /*compute the upper
                                                        noncentrality*/
coe = sqrt(nu);
```

```
NCt_lower=lamda_lower*coe;        /*Lower limit of noncentral ES*/
NCt_upper=lamda_upper*coe;        /*Upper limit of noncentral ES*/
Width= NCt_upper- NCt_lower;
TITLE "The confidence interval based on the noncentral method";
TITLE2 "Coefficient=&contrast Confidence level=&nominal";
PRINT NCt_lower NCt_upper Width;
QUIT;
%MEND;
data a;
input group y @@;
cards;
1  3 2 5 3 4 4 4
1  5 2 6 3 5 4 6
1  6 2 5 3 4 4 3
1  5 2 4 3 3 4 3
1  6 2 3 3 2 4 1
1  7 2 4 3 3 4 3
1  8 2 3 3 4 4 2
1 10 2 4 3 3 4 2
;
run;
%cinoncentral(a,.95,{.5 .5 -.5 -.5});
```

Appendix B

SAS® Macro “cibonett”

```
*************************************************************************;
*   format for data set: first variable = group, second variable = scores  ;
*   data = data set name                                                   ;
*   nominal = nominal confidence level                                     ;
*   contrast = a coefficient for each group, e.g., {.5 .5 -.5 -.5}         ;
*************************************************************************;

%macro cibonett(data,nominal,contrast);
TITLE "Bonett &nominal confidence interval";
proc IML;
use &data;
read all into datain;
groups = UNIQUE(datain[,1]);        *get group information of the data;
ngroups = NCOL(groups);             *get number of groups of the data;
CALL SYMPUTX("n_groups",ngroups);   *set the macro variable n_groups to be the
                                     number of groups;
%DO i = 1 %TO &n_groups;            *loop for groups;
group&i = datain[LOC(datain[,1]=&i),2]; /*obtain all the scores for each group*/
mu&i = mean(group&i);
var&i = var(group&i);
n&i = nrow(group&i);
%END;
mu=mu1;
var=var1;
n=n1;
```

```
%DO i=2 %TO &n_groups;
mu = mu//mu&i;
var = var//var&i;
n = n//n&i;
contrast = t(&contrast);
%END;
df=n-1;
delta_bonett=sum(mu#contrast)/sqrt(mean(var));
k = ngroups;
v1=(delta_bonett**2/(k**2*(mean(var))**2));
v2=sum((var##2)/(2*df));
v3=sum(((contrast##2)#var)/df)/mean(var);
var_delta_bonett=v1*v2+v3;
bonett_upper = delta_bonett + PROBIT(1-(1-&nominal)/2)*SQRT(var_delta_bonett);
bonett_lower = delta_bonett - PROBIT(1-(1-&nominal)/2)*SQRT(var_delta_bonett);
width=bonett_upper-bonett_lower;
PRINT delta_bonett bonett_lower bonett_upper width;
quit;
%mend;
data a;
input group y @@;
cards;
1  3 2 5 3 4 4 4
1  5 2 6 3 5 4 6
1  6 2 5 3 4 4 3
1  5 2 4 3 3 4 3
1  6 2 3 3 2 4 1
1  7 2 4 3 3 4 3
1  8 2 3 3 4 4 2
1 10 2 4 3 3 4 2
;
run;

%cibonett(a,.95,{.5 .5 -.5 -.5});
```

Appendix C

SAS® Macro “cibca”

```
*************************************************************************;
*   format for data set: first variable = group, second variable = scores  ;
*   data = data set name                                                   ;
*   b = the number of bootstrap sample                                     ;
*   con = a coefficient for each group, e.g., {.5 .5 -.5 -.5}              ;
*   nominal = nominal confidence level                                     ;
*************************************************************************;
%MACRO cibca(data=,b=,con=,nominal=);
/*************************************************************************/
/*This section of IML do the bootstrap resampling with B replications and     */
/*save the samples into zboots&i data sets (&i = 1 to number of groups)       */
/*It also calculate the delta_bonett for the original data set                */
/*************************************************************************/
```

```
PROC IML;
USE &data;
READ ALL INTO datain;
groups = UNIQUE(datain[,1]);         *get group information of the data;
ngroups = NCOL(groups);              *get number of groups of the data;
CALL SYMPUTX("n_groups",ngroups);    *set the macro variable n_groups to be
                                      the number of groups;
%DO i = 1 %TO &n_groups;             *loop for groups;
group&i = datain[LOC(datain[,1]=&i),2]; *obtain all the scores for each group;
mu&i = mean(group&i);
var&i = var(group&i);
group_n=NROW(group&i);
CALL SYMPUTX("n",group_n);
z&i = j(1,2,.);        *obtain the b times bootstrap sample for each group;
%DO m = 1 %TO &b;      *loop for the bootstrap samples;
       y = group&i;   /*This part of code (bootstrap) is adapted from     */
       z = j(&n, 1, 0) ; /*http://www.biostat.umn.edu/~john-c/5421/notes.016b*/
       ite = J(&n,1,&m); /*Identify the nth bootstrap sample */
              do j = 1 to &n ;
                     yrandindex = 1 + int(&n * ranuni(-1)) ;
                     z[j] = y[yrandindex] ;
              end ;
                           z = ite||z;
       z&i=z&i//z ;
%END;
CREATE zboots&i FROM z&i;    /*Put matrices to data sets the first column
                              is the index for boot sample*/;
APPEND FROM z&i;             *the second column is the boot sample data;
                             *Save the matrices from IML to SAS data set;
%END;
mu=mu1;         *do the delta_bonett for original data; /*for jackknife*/
var=var1;
%DO i=2 %TO &n_groups;
mu = mu||mu&i;
var = var||var&i;
%END;
o_delta_bonett=j(&b,1,(mu#&con)[,+]/sqrt((var[,+])/&n_groups));
       CREATE origbonett FROM o_delta_bonett ;
       APPEND FROM o_delta_bonett;  /* save the delta_bonett for original data
                                       to origbonett data*/
QUIT;

/*************************************************************************/
/* Below loop calculate the means and variances by each
   boot index for each groups */
/*************************************************************************/
%DO i = 1 %TO &n_groups;
proc means data = zboots&i noprint; /*Compute the group means and vars for each
                                     bootstrap sample*/
class COL1;
var COL2;
output out=meanvar&i mean=mu var=sigmasq;
run;
data meanvar&i;set meanvar&i;       *delete unused information;
```

```
if _type_=1;
drop _type_ _freq_;
run;
%END;
/*************************************************************************/
/*Below IML calculate the delta_bonett for the bootstrap data set*/
/*************************************************************************/
PROC IML;
%DO i = 1 %TO &n_groups;
USE meanvar&i;
READ ALL INTO mv&i;
%END;
mu = mv1[,2];                  *assign the means of first group to mu;
var = mv1[,3];                 *assign the vars of first group to var;
%DO i = 2 %TO &n_groups;      /*this loop add means and vars of other groups
                                to mu and var*/
mu = mu||mv&i[,2];
var = var||mv&i[,3];
%END;
con=repeat(&con,&b,1);         *make the contrast to a matrix for calculation;
delta_bonett=mv1[,1]||(mu#&con)[,+]/sqrt((var[,+])/&n_groups);
                              /*calculate delta_bonett for each bootstrap sample
                                and then add the bootstrap index to the first
                                column for later use*/;
CREATE delta_bonett FROM delta_bonett ;
APPEND FROM delta_bonett;
QUIT;
/*COMPUTE BIAS*/
data bonett           /* data set containing bootstrap values */
bias (keep=bias);     /* data set containing bias correction value */
merge delta_bonett(rename=(COL1=sample COL2=delta_bonett))
origbonett(rename=(COL1=origbonett)) end=eof;
if delta_bonett lt origbonett then lessthan=1;     /*flag if bootstrap sample
                                                     gives lower */
else lessthan=0;      /*value than original sample */
retain nless 0;       /*retain variable nless with starting value 0,
                        the second value of nless will be 0 add to the
                        first value of lessthan*/
if sample gt 0 then nless=nless+lessthan; /* count samples with flag lessthan */
if sample ne 0 then output bonett; /* output only bootstrap sample statistics */
if eof then do; /* for the last value calculate: */
propless=nless/sample; /* 1. proportion of values below original estimate */
bias=probit(propless); /* 2. inverse normal of that proportion */
output bias;           /* 3. output only that record to new data set */
end;
run;
/*JACKKNIFING ACCELERATION*/
data origjack;        /* create a new data set which contains observation */
set &data end=eof;    /* numbers 1 to &nobs (no. obs in data set) */
obsnum=_n_;
if eof then call symput('nobs', put(obsnum, 2.));  /*assign the characterstring
                                                     _n_ to macro variable nobs"*/
run;
%macro jackdata;             /* use macro for %do processing utility */
```

```
data jackdata;
set
%do i=1 %to &nobs;           /* do loop to create all samples */
origjack (in = in&i
where=(obsnum ne &i))        /* remove a different value each time */
%end;;
%do i=1 %to &nobs;
if in&i then repeat=&i;      /* add repeat number for each sample */
%end;
run;
%mend;
%jackdata;
proc means data=jackdata noprint nway; /*Group means for each JACKKNIFE sample*/
class repeat group;
var Y;
output out=jackmeanvar mean=mu var=sigmasq;
run;
data jackmeanvar;set jackmeanvar;          /*delete unused information*/
if _type_=3;
drop _type_ _freq_;
run;
proc transpose data=jackmeanvar out=jackmeanvar let; /*data restructure*/
by repeat;
id group;
run;
data jackmean ; set jackmeanvar;
if _NAME_="mu";
drop repeat _NAME_;
run;
data jackvar ; set jackmeanvar; /*Group variance for each JACKKNIF sample*/
if _NAME_="sigmasq";
drop repeat _NAME_;
run;
/*************************************************************************/
/*Below IML calculate the delta_bonett for JACKKNIFING ACCELERATION*/
/*************************************************************************/
PROC IML;
USE jackmean;
READ ALL INTO jackmean;
USE jackvar;
READ ALL INTO jackvar;
n_jack=NROW(jackmean);
CALL SYMPUTX("n_j",n_jack);
con=repeat(&con,&n_j,1);
delta_bonett=(jackmean#&con)[,+]/sqrt((jackvar[,+])/&n_groups);
CREATE jackbonett FROM delta_bonett ;
APPEND FROM delta_bonett;
QUIT;
DATA jackbonett (rename=(COL1 = delta_bonett));
SET jackbonett;
PROC SQL NOPRINT;
select mean(delta_bonett)    /* put mean of jackknifed values
                                into macro variable */
into :mean_delta_bonett
```

```
from jackbonett;
quit;
data meanbonett;
set jackbonett;
cubed=(&mean_delta_bonett - delta_bonett)**3;      /* create cubed value of
                                                     difference */
squared=(&mean_delta_bonett - delta_bonett)**2;    /* create squared value of
                                                     difference */
run;
proc means data=meanbonett noprint;
output out=sumbonett
sum(cubed)=sumcube           /* find sum of cubed values */
sum(squared)=sumsquar;       /* find sum of squared values */
run;
data accel;
set sumbonett;
accel=sumcube / (6 * (sumsquar**1.5));      /* plug values into equation for */
keep accel;                                 /* the acceleration statistic */
run;
data ciends;
merge accel
bias;
part1=(bias + probit((1-&nominal)/2))  / (1 - (accel*(bias + probit((1-
       &nominal)/2))));
part2=(bias + probit(1-(1-&nominal)/2))  / (1 - (accel*(bias + probit(1-(1-
       &nominal)/2))));
alpha1=probnorm(bias + part1);
alpha2=probnorm(bias + part2);
n1=alpha1*&b;
n2=alpha2*&b;
if n1 < 1 then n1 = 1;
call symput('n1', put(floor(n1), 5.)); /* Create macro variables with values */
call symput('n2', put(ceil(n2), 5.));  /* of N1 and N2 for later use */
run;
proc sort
data=bonett;
by delta_bonett;
run;
data ci_bca;
set bonett end=eof;
retain conf_lo conf_hi width;
if _n_=&n1 then conf_lo=delta_bonett; /* select values for upper and lower */
if _n_=&n2 then conf_hi=delta_bonett; /* limits using N1 and N2 values */
if eof then output;
width=conf_hi-conf_lo;
run;
proc print data=ci_bca;
title "The Confidence interval based on BCa bootstrap method";
title2 "B=&b Coefficeint=&con Confidence level=&nominal";
var conf_lo conf_hi width;
run;
%MEND;
Data BCa;
INPUT group y @@;
```

```
DATALINES;
1  3 2 5 3 4 4 4
1  5 2 6 3 5 4 6
1  6 2 5 3 4 4 3
1  5 2 4 3 3 4 3
1  6 2 3 3 2 4 1
1  7 2 4 3 3 4 3
1  8 2 3 3 4 4 2
1 10 2 4 3 3 4 2
;
%cibca(data=Bca,b=1000,con={.5 .5 -.5 -.5},nominal=.95); /*con is the
                                                          coefficient
                                                          for contrast*/
```

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 105-120.

ISSN 1538 − 9472

The Impact of Continuity Violation on ANOVA and Alternative Methods

Björn Lantz
Chalmers University of Technology
Gothenburg, Sweden

The normality assumption behind ANOVA and other parametric methods implies that response variables are measured on continuous scales. A simulation approach is used to explore the impact of continuity violation on the performance of statistical methods commonly used by applied researchers to compare locations across several groups.

Keywords: Continuity violation, ANOVA, Brown-Forsythe, Welch, Kruskal-Wallis

Introduction

One of the standard research procedures to explore the effects of the violation of an assumption underlying a statistical method is to perform an experimental study using Monte Carlo simulation. The one-way ANOVA for comparing locations across three or more groups and alternative test procedures such as the Brown-Forsythe test, Welch test, and Kruskal-Wallis test have been subject to similar research since the 1970s (e.g., Glass et al., 1972; Bevan et al., 1974; Keselman et al., 1977), and continue to be studied today (e.g., Lantz, 2013; Cribbie et al., 2012; Cribbie et al., 2007). Some workers conclud the one-way ANOVA is relatively robust against violations of the homoscedasticity assumption as well as against violations of the normality assumption. However, textbooks in statistics (e.g., Lomax and Hahs-Vaughn, 2007; Ryan, 2007) often recommend the Brown-Forsythe and Welch tests when the data are characterised by apparent heteroscedasticity, particularly at unequal sample sizes, or the Kruskal-Wallis test when the data are clearly not mound-shaped.

Most research regarding the normality assumption on which the ANOVA relies focuses on continuous distributions that differ from the normal in terms of shape, skewness, or kurtosis (e.g., Ito, 1980; Khan and Rayner, 2003). The fact that the underlying distribution is assumed to be normal does not, however, imply

Dr. Lantz is an Associate Professor in the Department of Technology Management and Economics, Chalmers University of Technology. Email him at: bjorn.lantz@chalmers.se.

only a mound shape, zero skewness, and zero excess kurtosis; it also requires that the data be continuous by nature. In applied research, data subject to statistical analyses have often been collected using discrete scales. Assume, for example, that the subjects participating in a psychological experiment perform a certain task four times, and that the number of successful trials is recorded for each subject. In this case, an arbitrarily chosen subject will have zero, one, two, three, or four successful trials. Although means and standard deviations can be used to describe the locations of different groups of subjects in cases like this, the one-way ANOVA and parametric alternatives like the Brown-Forsythe test and the Welch test are, at least technically, invalidated as methodologies to compare means across groups. This is because the dependent variable is assumed to be continuous even though it actually is discretely distributed, with only a small number of possible values.

The impact of the relative violation of the continuity assumption emerges more strikingly when there are fewer possible values that the variable can take. Krieg (1999) derived equations for calculating the bias induced by coarse measurement scales, and showed that the bias is reduced as the number of scale points increases. Hence, one would assume that statistical comparisons of locations across groups should be relatively unproblematic even if data are discrete as long as the number of possible variable values is large. In contrast, it might be a problem when the number of possible variable values is small, or when the violation of continuity is more severe. However, explicit analyses on the violation of continuity are scarce in the literature, and most of the research in this area seems to be related to the scale coarseness issue (Symonds, 1924) rather than to continuity violation. Scale coarseness refers to the fact that Likert-type and similar ordinal-level scales are collapsed into discrete scale points to simplify the data collection process, even though the underlying constructs are assumed to be continuous. When respondents are faced with a scale that does not have a sufficient number of response options, information loss will occur. Continuity violation and scale coarseness are obviously related phenomena, but scale coarseness (see Symonds, 1924) is an issue primarily related to data collection, whereas violation of continuity (see Bevan et al., 1974) is an issue strictly related to data analysis.

Although there seems to be little research on how continuity violation affects the statistical methods commonly used to compare locations across groups, nevertheless some results can be found in the literature. Bevan et al. (1974) considered the appropriateness of ANOVA techniques when the response variable was discretely distributed and able to take three, five, or seven different values.

Their results suggested that the ANOVA was relatively robust to continuity violations with respect to Type I errors. However, Bevan et al. (1974) did not examine how power or alternative methods were affected by continuity violation. Gregoire and Driver (1987) tested the performance of selected parametric (including the F test) and nonparametric tests of location on the basis of sampling results from simulated Likert-type data and concluded that there was no clear-cut superiority for either type of test. It should be noted that their aim was to compare the methods rather than to explore the impact of scale discreteness. Rasmussen (1988) extended (and corrected) the analysis by Gregoire and Driver (1987), and demonstrated that the Type I and Type II error rates were not seriously compromised by the use of discrete data.

The impact of continuity violation on the significance and power of statistical methods commonly used to compare locations across several groups is explored in the one-way ANOVA layout and its robust alternatives, the Brown-Forsythe test, the Welch test, and the non-parametric Kruskal-Wallis test. The one-way ANOVA is based on the idea that the true means in groups are more likely to be equal if the variation between the groups is small compared to the variation within the groups. The Brown-Forsythe and Welch tests are considered robust compared with ANOVA, because their definitions of variation within groups are based on the relationships between the different sample sizes in the different groups, as opposed to a simple pooled variance estimate, which means that they become less sensitive to heteroscedasticity (see, e.g., Tomarken & Serlin, 1986). The Kruskal-Wallis test is considered robust because it is based on ranks rather than actual values, which means that the underlying distribution does not matter so long as the observed values can be ranked.

Methodology

By definition, there is no discrete equivalent with only a few steps to the normal distribution, because a normally distributed random variable is unrestricted upward as well as downward, and can therefore take extreme values. Hence, it is technically impossible to make an exact evaluation of the impact of continuity violation on statistical methods that rest on the normality assumption. The best approximation compares results from a mound-shaped discrete distribution where the number of steps can be varied with results where the normality assumption holds, means and variances being equal. The binomial distribution is one such mound-shaped discrete distribution that exists for any number of steps, and it approaches the normal distribution when the number of steps becomes large

(Aczel and Sounderpandian, 2009). For example, Figure 1 displays the probability density function for the normal distribution with $\mu = 2$ and $\sigma^2 = 1$ and for the probability distribution for the binomial distribution with five possible outcomes, $\mu = 2$ and $\sigma^2 = 1$. Therefore, the binomial distribution is used in this study as an approximation of a continuity-violated normal distribution.

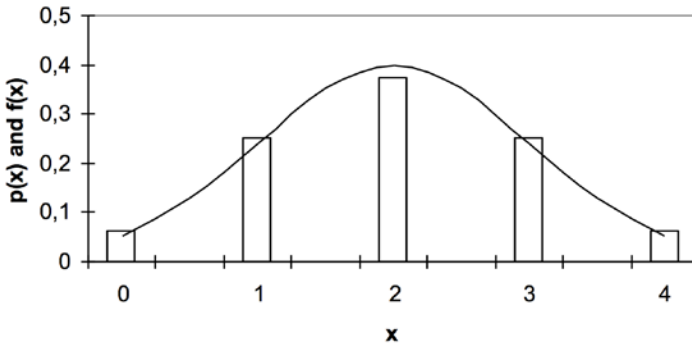


Figure 1. A binomial distribution and its corresponding normal distribution

An experimental design with three populations and four different combinations of small (defined as 5 observations) and large (defined as 25 observations) sample sizes was used. Discrete scales based on binomial distributions with two, three, four, five, and seven steps were used in each case. For each combination, the proportion of significant ANOVA, Brown-Forsythe, Welch, and Kruskal-Wallis (adjusted for ties) tests was compared to the proportion of significant tests when data was simulated from normal distributions with identical means and variances.

For each combination of sample sizes, test procedure, and number of steps, five different effect sizes were used. Table 1 shows the manner in which the values for the parameter p in the binomial distributions were varied for different values of n to achieve a suitable range of effect sizes (see Cohen, 1992), ranging from no effect ($f = 0.00$) to a very large effect ($f = 0.65$). For any individual combination of values of p and n, the distribution mean and variance could easily be calculated in order to obtain the corresponding normal distribution, because the mean is defined as np and the variance as $np(1-p)$ for the binomial distribution (Aczel and Sounderpandian, 2009). For example, with five steps ($n = 4$) and $f = 0.25$, $p_1 = 0.424$, $p_2 = 0.500$, and $p_3 = 0.576$ because the mean and the variance then become 2.12 and 1.22 for group 1, 2.50 and 1.25 for group 2, and 2.88 and 1.22 for group 3, respectively, corresponding to the medium effect size $f = 0.25$. Hence, the simulated impact of continuity violation in this case is based on a

comparison between the normal distributed random variables $X_1 \sim N(2.12, 1.22)$, $X_2 \sim N(2.50, 1.25)$, and $X_3 \sim N(2.88, 1.22)$ and the binomial distributed random variables $Y_1 \sim B(4, 0.424)$, $Y_2 \sim B(4, 0.500)$, and $Y_3 \sim B(4, 0.576)$.

Table 1. Values for the parameter *p* in the binomial distributions

		Effect size (Cohen's *f*)				
Steps	**Group**	***0.000***	***0.100***	***0.250***	***0.400***	***0.650***
2	1	0.500	0.439	0.351	0.273	0.166
	2	0.500	0.500	0.500	0.500	0.500
	3	0.500	0.561	0.649	0.727	0.834
3	1	0.500	0.457	0.394	0.334	0.245
	2	0.500	0.500	0.500	0.500	0.500
	3	0.500	0.543	0.607	0.667	0.756
4	1	0.500	0.465	0.413	0.346	0.285
	2	0.500	0.500	0.500	0.500	0.500
	3	0.500	0.535	0.587	0.654	0.715
5	1	0.500	0.469	0.424	0.380	0.311
	2	0.500	0.500	0.500	0.500	0.500
	3	0.500	0.531	0.576	0.620	0.689
7	1	0.500	0.475	0.438	0.401	0.343
	2	0.500	0.500	0.500	0.500	0.500
	3	0.500	0.525	0.562	0.599	0.657

For each combination of distribution (normal and binomial), sample sizes (25/25/25, 5/5/5, 5/5/25, and 5/25/25), test procedure (ANOVA, Brown-Forsythe, Welch, and Kruskal-Wallis), number of steps (two, three, four, five, and seven), and size of effect (no, small, medium, large, and very large), 50,000 hypothesis tests based on simulated random numbers were conducted, where the null hypothesis, corresponding to no difference between the locations of the populations, was challenged at an alpha level of 0.05 in all cases. Hence, 40,000,000 tests of simulated data were performed in the study. All simulations and analytical procedures were conducted using Microsoft Excel 2010.

Results

Table 2 displays the number of significant tests where the discrete scale has two steps. For a better understanding of the reliability of the statistics presented in this section, it should be noted that the standard error of a sample proportion at a sample size of 50,000 is about 0.002 when the proportion is 0.5, and it decreases to about 0.001 when the proportion is 0.05 or 0.95. When one distribution is characterised by a significantly larger proportion of significant tests than the other for a given combination of effect size, sample sizes, and test method, this is indicated with an asterisk (*).

Table 2: Proportion of significant tests, mean value 0.5 (two steps)

		ANOVA		Brown-Forsythe		Welch		Kruskal-Wallis	
ES	**n1/n2/n3**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**
0	25,25,25	0.052	0.051	0.052	0.050	0.051	0.051	0.052	0.05
	5,5,5	0.058*	0.052	0.057*	0.041	0.000	0.039*	0.059*	0.045
	5,5,25	0.052	0.051	0.046	0.047	0.000	0.052*	0.052*	0.044
	5,25,25	0.052	0.051	0.070*	0.053	0.011	0.056*	0.044	0.047
0.1	25,25,25	0.112*	0.107	0.112*	0.106	0.110*	0.104	0.112*	0.102
	5,5,5	0.071*	0.061	0.069*	0.049	0.000	0.046*	0.071*	0.052
	5,5,25	0.074*	0.070	0.061	0.062	0.001	0.067*	0.074*	0.060
	5,25,25	0.077*	0.072	0.092*	0.074	0.020	0.075*	0.068	0.066
0.25	25,25,25	0.468	0.466	0.468	0.465	0.463	0.466	0.468*	0.446
	5,5,5	0.131*	0.111	0.128*	0.091	0.000	0.088*	0.132*	0.098
	5,5,25	0.201	0.196	0.145	0.151*	0.009	0.152*	0.201*	0.170
	5,25,25	0.235*	0.223	0.215	0.215	0.079	0.199*	0.218*	0.207
0.4	25,25,25	0.858	0.879*	0.858	0.878*	0.855	0.889*	0.858	0.870*
	5,5,5	0.248*	0.214	0.244*	0.180	0.000	0.177*	0.251*	0.190
	5,5,25	0.447	0.455	0.308	0.321*	0.041	0.327*	0.445*	0.400
	5,25,25	0.508	0.506	0.453	0.481*	0.218	0.459*	0.491	0.484
0.65	25,25,25	0.999	1.000	0.999	1.000	0.978	1.000*	0.999	1.000
	5,5,5	0.525	0.519	0.521*	0.456	0.000	0.500*	0.535*	0.488
	5,5,25	0.855	0.907*	0.648	0.694*	0.147	0.760*	0.850	0.868*
	5,25,25	0.894	0.921*	0.845	0.910*	0.422	0.910*	0.889	0.919*

As a scale with two steps has the greatest degree of continuity violation, one would expect the most differences between the discrete binomial and continuous normal cases. When all sample sizes are small, the ANOVA becomes more powerful as a result of scale discreteness (i.e. the probability of avoiding a Type II error is often higher when data are discrete than when they are continuous), but at the cost of an elevated probability of a Type I error. For some combinations of

effect sizes and unequal sample sizes, the ANOVA becomes more powerful due to scale discreteness without an elevated probability of a Type I error. When all sample sizes are large, it becomes less powerful when the effect size is large, but more powerful when the effect size is small.

The Brown-Forsythe test becomes more powerful due to scale discreteness when all sample sizes are small and, for small and medium effect sizes, when exactly one sample size is small, but in both cases at the cost of an elevated probability of a Type I error. When exactly one sample size is large, it becomes less powerful for medium and larger effect sizes.

The Welch test algorithm does not work satisfactorily for coinciding dichotomous distributions when at least one sample size is small, which is the reason for the very low numbers for the discrete scale in those cases. Note, however, that for large sample sizes, it becomes more powerful when the effect size is small, but less powerful when the effect size is large.

The Kruskal-Wallis test becomes more powerful as a result of scale discreteness when at most one sample size is large, but in both cases at the cost of an elevated probability of a Type I error. For small and medium effect sizes, it becomes more powerful when exactly one sample size is small. For large sample sizes, however, it becomes more powerful at small and medium effect sizes but less powerful at large effect sizes.

Finally, note that there are no significant differences in performance between the four methods when they are used to analyse data on a discrete scale with two steps as long as the sample sizes are large; the only exception is that the Welch test performs less well when the effect size is very large.

Table 3 displays the number of significant tests where the discrete scale has three steps. Here, the ANOVA shows no significant difference in performance due to scale discreteness, with the exception that it becomes more powerful when all sample sizes are small and the effect size is large. The Brown-Forsythe test exhibits elevated power when at least one sample size is small, but again, at the cost of an elevated probability of a Type I error. The Welch test displays the opposite reaction: it becomes less powerful when at least one sample size is small, but with a reduced probability of a Type I error. The Kruskal-Wallis test behaves erratically for some sample size combinations, and becomes less powerful at some effect sizes but more powerful at others. However, there is no significant change in the probability of a Type I error for any combination of sample sizes. Finally, note that there are no significant differences in performance between the four methods when they are used to analyse data on a discrete scale with three steps as long as the sample sizes are large.

Table 3: Proportion of significant tests, mean value 1.0 (three steps)

		ANOVA		Brown-Forsythe		Welch		Kruskal-Wallis	
ES	**n1/n2/n3**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**
0	25,25,25	0.051	0.05	0.051	0.049	0.052*	0.049	0.049	0.048
	5,5,5	0.054	0.052	0.049*	0.041	0.033	0.041*	0.044	0.046
	5,5,25	0.050	0.050	0.053*	0.047	0.037	0.054*	0.046	0.044
	5,25,25	0.050	0.051	0.056*	0.053	0.044	0.057*	0.046	0.046
0.1	25,25,25	0.108	0.108	0.108	0.107	0.108	0.105	0.104	0.103
	5,5,5	0.060	0.062	0.055*	0.050	0.038	0.049*	0.049	0.054*
	5,5,25	0.069	0.071	0.067*	0.062	0.049	0.067*	0.062	0.062
	5,25,25	0.073	0.074	0.080*	0.074	0.063	0.075*	0.069	0.068
0.25	25,25,25	0.461	0.46	0.460	0.459	0.457	0.455	0.453*	0.437
	5,5,5	0.113	0.110	0.106*	0.091	0.076	0.088*	0.094	0.096
	5,5,25	0.194	0.190	0.158*	0.149	0.127	0.152*	0.174*	0.168
	5,25,25	0.226	0.224	0.221*	0.212	0.189	0.197*	0.214*	0.207
0.4	25,25,25	0.865	0.875	0.864	0.874	0.859	0.877*	0.859	0.861
	5,5,5	0.222*	0.211	0.209*	0.178	0.156	0.172*	0.191	0.186
	5,5,25	0.438	0.435	0.333*	0.324	0.294	0.319*	0.402*	0.385
	5,25,25	0.509	0.508	0.474	0.469	0.434	0.444*	0.491*	0.479
0.65	25,25,25	0.999	1.000	0.999	1.000	0.999	1.000	0.999	0.999
	5,5,5	0.496	0.501	0.474*	0.446	0.336	0.442*	0.444	0.458*
	5,5,25	0.848	0.873*	0.686	0.706*	0.614	0.711*	0.812	0.827*
	5,25,25	0.905	0.918*	0.852	0.886*	0.791	0.880*	0.894	0.905

Table 4 displays the number of significant tests where the discrete scale has four steps. In this case, the ANOVA shows no significant difference in performance due to scale discreteness, except that it becomes powerful when at most one sample size is large and the effect size is very large. The Brown-Forsythe test becomes more powerful when all sample sizes are small, but less powerful at unequal sample sizes when the effect size is very large. The Welch test performs somewhat erratically, as it exhibits reduced power when sample sizes are unequal, but increased power when all sample sizes are small and the effect size is medium or large. The Kruskal-Wallis test also behaves erratically: it becomes too conservative when all sample sizes are small, which reduces power. In contrast, it becomes more powerful at the medium effect size when all sample sizes are large and at unequal sample sizes when the effect size is medium or large. As in the previous cases, note that there are no significant differences in performance between the four methods when they are used to analyse data on a discrete scale with four steps as long as the sample sizes are large.

Table 4: Proportion of significant tests, mean value 1.5 (four steps)

		ANOVA		Brown-Forsythe		Welch		Kruskal-Wallis	
ES	**n1/n2/n3**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**
0	25,25,25	0.050	0.052	0.049	0.052	0.050	0.051	0.048	0.049
	5,5,5	0.049	0.052	0.044	0.042	0.040	0.040	0.040	0.045*
	5,5,25	0.050	0.050	0.051*	0.047	0.053	0.054	0.045	0.044
	5,25,25	0.050	0.052	0.055	0.054	0.054	0.056	0.047	0.048
0.1	25,25,25	0.111	0.110	0.110	0.109	0.110	0.107	0.107	0.105
	5,5,5	0.059	0.062	0.053*	0.049	0.047	0.047	0.049	0.054*
	5,5,25	0.072	0.071	0.065	0.062	0.066	0.068	0.063	0.061
	5,25,25	0.075	0.074	0.079	0.076	0.072	0.076*	0.070	0.070
0.25	25,25,25	0.459	0.457	0.458	0.456	0.455	0.451	0.449*	0.434
	5,5,5	0.109	0.111	0.100*	0.091	0.092*	0.087	0.093	0.098*
	5,5,25	0.191	0.189	0.151	0.149	0.132	0.150*	0.169	0.165
	5,25,25	0.227	0.224	0.215	0.21	0.181	0.194*	0.213*	0.204
0.4	25,25,25	0.933	0.939	0.933	0.938	0.930	0.940	0.928	0.930
	5,5,5	0.264	0.257	0.244*	0.222	0.221*	0.210	0.228	0.228
	5,5,25	0.530	0.525	0.403	0.40	0.352	0.389*	0.479*	0.470
	5,25,25	0.618	0.615	0.563	0.565	0.512	0.540*	0.596*	0.584
0.65	25,25,25	0.999	0.999	0.999	0.999	0.999	1.000	0.999	0.999
	5,5,5	0.492	0.504*	0.465*	0.452	0.425	0.435*	0.441	0.458*
	5,5,25	0.852	0.864*	0.703	0.714*	0.651	0.707*	0.808	0.816
	5,25,25	0.913	0.918	0.862	0.879*	0.835	0.872*	0.900	0.903

Table 5 displays the number of significant tests where the discrete scale has five steps. The ANOVA displays a similar pattern as with four steps; there is no significant difference in performance due to scale discreteness, except that it becomes powerful when exactly one sample size is large and the effect size is very large. The Brown-Forsythe test also shows a similar pattern (as in the previous case), becoming more powerful when all sample sizes are small and the effect size is at least medium, but less powerful at unequal sample sizes when the effect size is very large. The performance of the Welch test, however, behaves somewhat differently when the number of steps is increased from four to five. It becomes more conservative when at least one sample size is small, which reduces its power when the effect size is small. The effect disappears when the effect size is medium or large, but returns when it is very large. The Kruskal-Wallis test continues to behave erratically along the same pattern as with four steps. Finally, under a medium effect size, the Kruskal-Wallis test has significantly less power than the other three methods even if all sample sizes are large.

Table 5: Proportion of significant tests, mean value 2.0 (five steps)

		ANOVA		Brown-Forsythe		Welch		Kruskal-Wallis	
ES	**n1/n2/n3**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**
0	25,25,25	0.050	0.050	0.050	0.049	0.049	0.049	0.047	0.047
	5,5,5	0.050	0.052	0.044	0.042	0.037	0.041*	0.041	0.045*
	5,5,25	0.050	0.050	0.049	0.047	0.047	0.053*	0.044	0.044
	5,25,25	0.049	0.049	0.053	0.053	0.053	0.057*	0.046	0.047
0.1	25,25,25	0.111	0.110	0.111	0.110	0.110	0.108	0.107	0.105
	5,5,5	0.059	0.062	0.052	0.050	0.045	0.048*	0.049	0.054*
	5,5,25	0.071	0.073	0.064	0.064	0.064	0.071*	0.062	0.064
	5,25,25	0.076	0.078	0.079	0.080	0.074	0.080*	0.071	0.072
0.25	25,25,25	0.463	0.462	0.463	0.461	0.458	0.456	0.450*	0.441
	5,5,5	0.110	0.108	0.099*	0.090	0.086	0.086	0.092	0.096*
	5,5,25	0.192	0.188	0.153	0.148	0.151	0.148	0.170*	0.165
	5,25,25	0.230	0.227	0.218	0.213	0.201	0.196	0.215*	0.207
0.4	25,25,25	0.864	0.869	0.864	0.869	0.860	0.867	0.854	0.852
	5,5,5	0.216	0.216	0.197*	0.183	0.173	0.174	0.185	0.189
	5,5,25	0.433	0.431	0.332	0.331	0.317	0.318	0.389	0.384
	5,25,25	0.515	0.511	0.466	0.468	0.437	0.436	0.491*	0.479
0.65	25,25,25	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	5,5,5	0.493	0.500	0.464*	0.448	0.415	0.428*	0.443	0.452*
	5,5,25	0.846	0.858*	0.706	0.718*	0.664	0.702*	0.803	0.811
	5,25,25	0.912	0.918	0.861	0.875*	0.843	0.869*	0.899	0.902

Table 6 displays the number of significant tests where the discrete scale has seven steps. The ANOVA now becomes more conservative when at most one sample size is large, and it has reduced power at the medium effect size when all sample sizes are small. Both the Brown-Forsythe test and the Welch test lose power at very large effect sizes when sample sizes are unequal, but become more powerful at the large effect size when all sample sizes are small. The Kruskal-Wallis test becomes too conservative and loses power when all sample sizes are small. It is also characterised by significantly less power than the other three methods when the effect size is small or medium.

Table 6: Proportion of significant tests, mean value 3.0 (seven steps)

		ANOVA		Brown-Forsythe		Welch		Kruskal-Wallis	
ES	**n1/n2/n3**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**	**Bin**	**Norm**
0	25,25,25	0.051	0.051	0.051	0.051	0.051	0.051	0.049	0.049
	5,5,5	0.048	0.052*	0.041	0.042	0.040	0.040	0.040	0.046*
	5,5,25	0.049	0.052*	0.047	0.048	0.051	0.054	0.043	0.045
	5,25,25	0.050	0.05	0.054	0.053	0.057	0.057	0.047	0.046
0.1	25,25,25	0.108	0.109	0.108	0.108	0.108	0.106	0.103	0.102
	5,5,5	0.059	0.058	0.050	0.047	0.048	0.046	0.048	0.051*
	5,5,25	0.072	0.069	0.063*	0.060	0.068	0.067	0.063	0.061
	5,25,25	0.076	0.076	0.077	0.075	0.076	0.075	0.071	0.068
0.25	25,25,25	0.457	0.459	0.456	0.458	0.451	0.452	0.440	0.436
	5,5,5	0.108	0.114*	0.094	0.094	0.088	0.088	0.091	0.099*
	5,5,25	0.188	0.188	0.152	0.149	0.152	0.148	0.166	0.166
	5,25,25	0.226	0.223	0.213	0.21	0.198	0.196	0.211	0.205
0.4	25,25,25	0.869	0.870	0.869	0.869	0.863	0.866	0.856	0.853
	5,5,5	0.215	0.213	0.191*	0.181	0.177*	0.170	0.184	0.187
	5,5,25	0.427	0.429	0.331	0.326	0.315	0.317	0.380	0.379
	5,25,25	0.519	0.514	0.472	0.465	0.442	0.437	0.49	0.482
0.65	25,25,25	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	5,5,5	0.491	0.499	0.453	0.447	0.420	0.416	0.437	0.449*
	5,5,25	0.842	0.852	0.708	0.718	0.673	0.693*	0.794	0.806*
	5,25,25	0.913	0.921	0.860	0.873*	0.848	0.867*	0.896	0.902

Table 7 provides a qualitative summary of the simulation results. Note that cases where continuity violation has no or negligible impact on the probability of a Type I error (α) or power ($1 - \beta$) are not explicitly discussed. For the ANOVA, scale discreteness is considered to have a marked impact on power because the number of differences between the normal distribution and the binomial distribution that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases. Sample sizes are also considered to have a marked impact on power, because power is reduced in several cases where sample sizes are unequal, but not when they are equal. However, continuity violation was not found to have a marked impact on the probability of a Type I error in any of the examined aspects, which is in line with previous research (Bevan et al., 1974).

Table 7: Impact of continuity violation – summary of simulation results

	ANOVA		**Brown-Forsythe**		**Welch**		**Kruskal-Wallis**	
Explanatory variable	*α*	1–*β*	*α*	1–*β*	*α*	1–*β*	*α*	1–*β*
Scale discreteness	Negligible impact	Marked impact	Marked impact	Marked impact	Marked impact	Marked impact	Marked impact	Marked impact
Effect size	n/a	Negligible impact	n/a	Negligible impact	n/a	Marked impact	n/a	Negligible impact
Sample sizes	Negligible impact	Marked impact	Negligible impact	Marked impact	Marked impact	Marked impact	Marked impact	Marked impact

For the Brown-Forsythe test, scale discreteness is considered to have a marked impact on the probability of a Type I error because the significant differences between the normal and binomial distributions that can be seen when the discrete scale has only a few steps and at least one sample size is small seem to decrease when the number of steps increases. Furthermore, scale discreteness is considered to have a marked impact on power because the number of differences between the normal and binomial distributions that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases. Sample sizes are also considered to have a marked impact on power because power is increased in several cases where at least one sample size is small.

For the Welch test, scale discreteness is considered to have a marked impact on the probability of a Type I error because the significant differences that can be seen between the normal and binomial distributions when the discrete scale has only at a few steps seem to decrease when the number of steps increase. Sample sizes are also considered to have a marked impact on the probability of a Type I error because this probability is consistently different in several cases where at least one sample size is small, but not when all sample sizes are large. Furthermore, scale discreteness is considered to have a marked impact on power because the number of differences between the normal distribution and the binomial distribution that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases. Effect size is also considered to have a marked impact on power, particularly in combination with scale discreteness, because the number of observable differences between the normal and binomial distributions tends to decrease faster at the medium and large effect sizes than at the small and very large effect sizes. In addition, sample

sizes are also considered to have a marked impact on power because the presence of unequal sample sizes seems to reduce power in general.

Finally, for the Kruskal-Wallis test, scale discreteness is considered to have a marked impact on the probability of a Type I error because the significant differences that can be seen between the normal and binomial distributions when the discrete scale has only a few steps, and all sample sizes are small, seem to be reversed when the number of steps increases. Sample sizes are also considered to have a marked impact on the probability of a Type I error because this probability is consistently different when all sample sizes are small, but not otherwise. Furthermore, scale discreteness is considered to have a marked impact on power because the number of differences between the normal and binomial distributions that can be seen when the discrete scale has only two steps seems to decrease when the number of steps increases, although the major difference occurs between two and three steps. Sample sizes are also considered to have a marked impact on power because power is changed in several cases where at least one sample size is small.

Conclusion

Violation of continuity affects the performance of four statistical methods that are commonly used to compare locations across several groups. A dichotomous scale changes the probability of a Type I error for methods in all cases when all sample sizes are small and in many other cases when at least one sample size is small. However, the effect seems to decline as the number of scale points is increased, which is in line with theory (Krieg, 1999) and with similar published simulation results (e.g., Bevan et al., 1974). The probability of a Type II error also seems to decline as the number of scale points is increased, although the pattern is different for different methods and sample size combinations.

This should not be seen as an argument in favour of a larger number of steps when, for example, Likert-type and similar discrete scales are used. Even a small number of steps may be too many for the respondent if comprehensible instructions and labelling of response alternatives are not included to enable the respondent to conceptualize and respond in spatial terms (Cox, 1980). Often, and for a variety of reasons, scales with only a few steps must be used during data collection processes, and the results in this study can help determine a suitable statistical procedure to compare locations across groups in such situations.

In summary, ANOVA seems to be the most robust alternative of the four procedures when scales are discrete, as the violation of continuity has relatively

little impact on its performance. The Brown-Forsythe test can become more powerful when scales are discrete and at least one sample size is small, but at the cost of an elevated probability of a Type I error. When all sample sizes are large and scales with at least three steps are used, neither the ANOVA nor the Brown-Forsythe test displays any significant sensitivity to continuity violation at any effect size level. Hence, these two tests can be used to make reliable analyses of discrete data in such situations. The Welch test can become less powerful when scales are discrete, in some cases even at large sample sizes. The Kruskal-Wallis test responds erratically to scale discreteness, particularly at unequal sample sizes, and has significantly less power than the other three methods when sample sizes are large.

Even though the impact of continuity violation on ANOVA and the three alternative methods examined here seem to be relatively small in most realistic situations (the most obvious exception is when the Welch test is used to analyse dichotomous data), applied researchers should consider the above results when using these statistical methods to analyse data collected with discrete scales. The main implications of this study can be summarised as follows:

- Collect data using continuous scales, if reasonable.
- Be aware that power can be reduced when discrete scales are used. The reduction in power becomes less pronounced when the number of scale points is increased, but in some situations, it remains significant for scales with up to seven points.
- Be aware that the actual probability of a Type I error may be affected when dichotomous scales are used if at least one sample size is small.
- Do not use the Welch test with dichotomous data.

Future research in this area should further explore the effects of data discreteness by combining continuity violation with, for example, heteroscedasticity. In general, the effects of concurrent violations can produce anomalous effects not observed in separate violations (see, e.g., Zimmerman, 1998). Other types of parametric methods should also be tested for their robustness against continuity violation.

References

Aczel, A. D. & Sounderpandian, J. (2009). *Complete Business Statistics*, New York: McGraw-Hill Irwin.

Bevan, M. F., Denton, J. Q., & Myers, J. L. (1974). The robustness of the *F* test to violations of continuity and form of treatment population. *British Journal of Mathematical and Statistical Psychology*, *27*, 199-204. doi: 10.1111/j.2044-8317.1974.tb00540.x

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi: 10.1037/0033-2909.112.1.155

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *27*, 407-422. doi: 10.2307/3150495

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology*, *65*, 56-73. doi: 10.1111/j.2044-8317.2011.02014.x

Cribbie, R. A., Wilcox, R. R., Bewell, C., & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, *6*, 117-132.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, *42*, 237-288. doi: 10.3102/00346543042003237

Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin, 101*, 159-165. doi: 10.1037/0033-2909.101.1.159

Ito, K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnainh, Ed., *Handbook of Statistics*, Vol. 1, Amsterdam, Holland.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977), An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, *30*: 213-221. doi: 10.1111/j.2044-8317.1977.tb00742.x

Khan, A., & Rayner, G. D. (2003). Robustness to non-normality of common tests for the many sample location problem. *Journal of Applied Mathematics and Decision Sciences*, *7*(4), 187-206. doi: 10.1155/S1173912603000178

Krieg, E. F. (1999). Biases induced by coarse measurement scales. *Educational and Psychological Measurement, 59,* 749-766. doi: 10.1177/00131649921970125

Lantz, B. (2013). The impact of sample non-normality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology, 66*, 224-244. doi: 10.1111/j.2044-8317.2012.02047.x

Lomax, R. G. & Hahs-Vaughn, D. L. (2007). *An Introduction to Statistical Concepts*. NJ: Lawrence Erlbaum Associates.

Rasmussen, J. L. (1988). Analysis of Likert-Scale Data: A Reinterpretation of Gregoire and Driver. *Psychological Bulletin*, *105*, 167-170. doi: 10.1037/0033-2909.105.1.167

Ryan, T. P. (2007). *Modern Experimental Design*. Hoboken, NJ: John Wiley & Sons.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456-461. doi: 10.1037/h0074469

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*, 90-99.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, *67*(1), 55-68. doi: 10.1080/00220979809598344

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 121-155.

ISSN 1538 − 9472

Generalized Modified Ratio Estimator for Estimation of Finite Population Mean

Jambulingam Subramani
Pondicherry University
Puducherry, India

A generalized modified ratio estimator is proposed for estimating the population mean using the known population parameters. It is shown that the simple random sampling without replacement sample mean, the usual ratio estimator, the linear regression estimator and all the existing modified ratio estimators are the particular cases of the proposed estimator. The bias and the mean squared error of the proposed estimator are derived and are compared with that of existing estimators. The conditions for which the proposed estimator performs better than the existing estimators are also derived. The performance of the proposed estimator is assessed with that of the existing estimators for certain natural populations

Keywords: Auxiliary variable, biases, natural population, mean squared error, parameters

Introduction

Consider a finite population $U = \{ U_1, U_2, \dots , U_N \}$ of N distinct and identifiable units. Let Y be a study variable with value Y_i measured on U_1, $i = 1, 2, 3, \dots , N$ giving a vector $Y = \{ Y_1, Y_2, \dots , Y_N \}$. The problem is to estimate the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ on the basis of a random sample selected from the population U. The simple random sample mean is the simplest estimator for estimating the population mean. If an auxiliary variable X, closely related to the study variable Y, is available then one can improve the performance of the estimator of the study variable by using the known values of the population parameters of the auxiliary variable. That is, when the population parameters of the auxiliary variable X such as population mean, coefficient of variation, coefficient of kurtosis, coefficient of skewness etc., are known, then a number of estimators available in the literature (such as ratio, product and linear regression

Dr. Subramani is Associate Professor and Head of the Department of Statistics. Email him at: drjsubramani@yahoo.co.in.

estimators and their modifications) perform better than the usual simple random sample mean under certain conditions. Among these estimators, many researchers have used the ratio estimator and its modifications for the estimation of the mean of the study variable (see for example Sisodia and Dwivedi (1981), Kadilar and Cingi (2006a, 2006b), Yan and Tian (2010) and Subramani and Kumarapandiyan (2012a, 2012c)). Before discussing further the existing estimators and the proposed estimators, the notations to be used in this article are described below:

N	Population size
n	Sample size
$f = n/N$	Sampling fraction
Y	Study variable
X	Auxiliary variable
$\bar{X}, \bar{Y}$	Population means
$\bar{x}, \bar{y}$	Sample means
S_x, S_y	Population standard deviations
C_x, C_y	Co-efficient of variations
ρ	Co-efficient of correlation between X and Y
β_1	Co-efficient of skewness of the auxiliary variable
β_2	Co-efficient of kurtosis of the auxiliary variable
M_d	Median of the auxiliary variable
$B(.)$	Bias of the estimator
$MSE(.)$	Mean squared error of the estimator
$\hat{\bar{Y}}_i\left(\hat{\bar{Y}}_{p_j}\right)$	ith existing (jth proposed) modified ratio estimator of $\bar{Y}$

In case of simple random sampling without replacement (SRSWOR), the sample mean $\bar{y}_{srs}$ is used to estimate population mean $\bar{Y}$, which is an unbiased estimator, and its variance is given below:

$$V\left(\bar{y}_{srs}\right) = \frac{\left(1-f\right)}{n} S_y^2 \tag{1}$$

The ratio estimator for estimating the population mean $\bar{Y}$ of the study variable Y is defined as:

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{R}\bar{X} \text{ where } \hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{y}{x} \tag{2}$$

The bias and mean squared error of the ratio estimator to the first degree of approximation are given below:

$$B\left(\hat{\bar{Y}}_R\right) = \frac{(1-f)}{n}\bar{Y}\left(C_x^2 - \rho C_x C_y\right) \tag{3}$$

$$MSE\left(\hat{\bar{Y}}_R\right) = \frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_x C_y\right) \tag{4}$$

The usual linear regression estimator together with its variance is given below:

$$\hat{\bar{Y}}_{lr} = \bar{y} + \beta\left(\bar{X} - \bar{x}\right) \tag{5}$$

$$V\left(\hat{\bar{Y}}_{lr}\right) = \frac{(1-f)}{n}S_y^2\left(1-\rho^2\right) \tag{6}$$

Sisodia and Dwivedi (1981) have suggested a modified ratio estimator using the co-efficient of variation of auxiliary variable X for estimating $\bar{Y}$. When the co-efficient of kurtosis of auxiliary variable X is known, Singh et al. (2004) has developed a modified ratio estimator. Singh and Tailor (2003) proposed another estimator for estimating $\bar{Y}$ when the population correlation co-efficient between X and Y is known. By using the population variance of auxiliary variable X, Singh (2003) proposed another modified ratio estimator for estimating population mean. More recently, Yan and Tian (2010) has suggested another modified ratio estimator using the co-efficient of skewness of the auxiliary variable X, and Subramani and Kumarapandiyan (2013a) suggested a new modified ratio estimator using known population median of auxiliary variable X.

Upadhyaya and Singh (1999) suggested another modified ratio estimator using the linear combination of co-efficient of variation and co-efficient of kurtosis. Singh (2003) used the linear combination of co-efficient of kurtosis and standard deviation and co-efficient of skewness and standard deviation for estimating the populations mean $\bar{Y}$. Motivated by Singh (2003), Yan and Tian (2010) used the linear combination of co-efficient of kurtosis and co-efficient of

skewness and co-efficient of variation and co-efficient of skewness. Subramani and Kumarapandiyan (2012a, 2012b, 2012c and 2013b) suggested modified ratio estimators using known median and co-efficient of kurtosis, median and co-efficient of skewness, median and co-efficient of variation and median and co-efficient of correlation.

More detailed discussion about the ratio estimator and its modification can be found in Abdia and Shahbaz (2006), Ahmad et al. (2009), Al-Jararha and Al-Haj Ebrahem (2012), Bhushan (2012), Cochran (1977), Dalabehera and Sahoo (1994), David and Sukhatme (1974), Goodman and Hartley (1958), Gupta and Shabbir (2008), Jhajj et al. (2006), Kadilar and Cingi (2003, 2004), Khoshnevisan et al. (2007), Koyuncu and Kadilar (2009), Kulkarni (1978), Murthy (1967), Naik and Gupta (1991), Olkin (1958), Pathak (1964), Perri (2007), Ray and Sahai (1980), Reddy (1973), Robinson (1987), Sen (1993), Shabbir and Yaab (2003), Sharma and Tailor (2010), Singh and Chaudhary (1986), Singh (2003), Singh and Espejo (2003), Singh and Agnihotri (2008), Singh and Solanki (2012), Singh and Tailor (2003, 2005), Singh et al. (2004, 2008), Sisodia and Dwivedi (1981) , Solanki et al. (2012), Srivenkataramana (1980), Tailor and Sharma (2009), Tin (1965), Upadhyaya and Singh (1999) and Yan and Tian (2010).

The following table contains all modified ratio estimators using known population parameters of the auxiliary variable in which some of the estimators are already suggested in the literature. The remaining estimators are introduced in this article:

Table 1. Modified Ratio estimators with the constant, the bias, and the mean squared errors.

Estimator	Constant θ_i	Bias – $B(.)$	Mean squared error MSE(.)
$\hat{\bar{Y}} = \bar{y}\left[\frac{\bar{X}+C_x}{\bar{x}+C_x}\right]$ *Sisodia and Dwivedi (1981)*	$\theta_1 = \frac{\bar{X}}{\bar{X}+C_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_1^2 C_x^2 - \theta_1 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_1^2 C_x^2 - 2\theta_1 \rho C_x C_y\right)$
$\hat{\bar{Y}}_2 = \bar{y}\left[\frac{\bar{X}+\beta_2}{\bar{x}+\beta_2}\right]$ *Singh et al. (2004)*	$\theta_2 = \frac{\bar{X}}{\bar{X}+\beta_2}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_2^2 C_x^2 - \theta_2 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_2^2 C_x^2 - 2\theta_2 \rho C_x C_y\right)$
$\hat{\bar{Y}}_3 = \bar{y}\left[\frac{\bar{X}+\beta_1}{\bar{x}+\beta_1}\right]$ *Yan and Tian (2010)*	$\theta_3 = \frac{\bar{X}}{\bar{X}+\beta_1}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_3^2 C_x^2 - \theta_3 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_3^2 C_x^2 - 2\theta_3 \rho C_x C_y\right)$
$\hat{\bar{Y}}_4 = \bar{y}\left[\frac{\bar{X}+\rho}{\bar{x}+\rho}\right]$ *Singh and Tailor (2003)*	$\theta_4 = \frac{\bar{X}}{\bar{X}+\rho}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_4^2 C_x^2 - \theta_4 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_4^2 C_x^2 - 2\theta_4 \rho C_x C_y\right)$

Table 1 Continued

Estimator	Constant θ_i	Bias – $B(.)$	Mean squared error MSE(.)
$\hat{\bar{Y}}_5 = \bar{y}\left[\frac{\bar{X}+S_x}{\bar{x}+S_x}\right]$ *Singh (2003)*	$\theta_5 = \frac{\bar{X}}{\bar{X}+S_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_5^2 C_x^2 - \theta_5 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_5^2 C_x^2 - 2\theta_5 \rho C_x C_y\right)$
$\hat{\bar{Y}}_6 = \bar{y}\left[\frac{\bar{X}+M_d}{\bar{x}+M_d}\right]$ *Subramani and Kumarapandiyan (2013a)*	$\theta_6 = \frac{\bar{X}}{\bar{X}+M_d}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_6^2 C_x^2 - \theta_6 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_6^2 C_x^2 - 2\theta_6 \rho C_x C_y\right)$
$\hat{\bar{Y}}_7 = \bar{y}\left[\frac{\beta_2\bar{X}+C_x}{\beta_2\bar{x}+C_x}\right]$ *Upadhyaya and Singh (1999)*	$\theta_7 = \frac{\beta_2\bar{X}}{\beta_2\bar{X}+C_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_7^2 C_x^2 - \theta_7 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_7^2 C_x^2 - 2\theta_7 \rho C_x C_y\right)$
$\hat{\bar{Y}}_8 = \bar{y}\left[\frac{C_x\bar{X}+\beta_2}{C_x\bar{x}+\beta_2}\right]$ *Upadhyaya and Singh (1999)*	$\theta_8 = \frac{C_x\bar{X}}{C_x\bar{X}+\beta_2}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_8^2 C_x^2 - \theta_8 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_8^2 C_x^2 - 2\theta_8 \rho C_x C_y\right)$
$\hat{\bar{Y}}_9 = \bar{y}\left[\frac{\beta_1\bar{X}+C_x}{\beta_1\bar{x}+C_x}\right]$	$\theta_9 = \frac{\beta_1\bar{X}}{\beta_1\bar{X}+C_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_9^2 C_x^2 - \theta_9 \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_9^2 C_x^2 - 2\theta_9 \rho C_x C_y\right)$
$\hat{\bar{Y}}_{10} = \bar{y}\left[\frac{C_x\bar{X}+\beta_1}{C_x\bar{x}+\beta_1}\right]$ *Yan and Tian (2010)*	$\theta_{10} = \frac{C_x\bar{X}}{C_x\bar{X}+\beta_1}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{10}^2 C_x^2 - \theta_{10} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{10}^2 C_x^2 - 2\theta_{10} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{11} = \bar{y}\left[\frac{\rho\bar{X}+C_x}{\rho\bar{x}+C_x}\right]$	$\theta_{11} = \frac{\rho\bar{X}}{\rho\bar{X}+C_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{11}^2 C_x^2 - \theta_{11} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{11}^2 C_x^2 - 2\theta_{11} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{12} = \bar{y}\left[\frac{C_x\bar{X}+\rho}{C_x\bar{x}+\rho}\right]$	$\theta_{12} = \frac{C_x\bar{X}}{C_x\bar{X}+\rho}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{12}^2 C_x^2 - \theta_{12} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{12}^2 C_x^2 - 2\theta_{12} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{13} = \bar{y}\left[\frac{S_x\bar{X}+C_x}{S_x\bar{x}+C_x}\right]$	$\theta_{13} = \frac{S_x\bar{X}}{S_x\bar{X}+C_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{13}^2 C_x^2 - \theta_{13} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{11}^2 C_x^2 - 2\theta_{11} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{14} = \bar{y}\left[\frac{C_x\bar{X}+S_x}{C_x\bar{x}+S_x}\right]$	$\theta_{14} = \frac{C_x\bar{X}}{C_x\bar{X}+S_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{14}^2 C_x^2 - \theta_{14} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{14}^2 C_x^2 - 2\theta_{14} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{15} = \bar{y}\left[\frac{M_d\bar{X}+C_x}{M_d\bar{x}+C_x}\right]$	$\theta_{15} = \frac{M_d\bar{X}}{M_d\bar{X}+C_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{15}^2 C_x^2 - \theta_{15} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{15}^2 C_x^2 - 2\theta_{15} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{16} = \bar{y}\left[\frac{C_x\bar{X}+M_d}{C_x\bar{x}+M_d}\right]$ *Subramani and Kumarapandiyan (2012c)*	$\theta_{16} = \frac{C_x\bar{X}}{C_x\bar{X}+M_d}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{16}^2 C_x^2 - \theta_{16} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{16}^2 C_x^2 - 2\theta_{16} \rho C_x C_y\right)$
$\hat{\bar{Y}}_{17} = \bar{y}\left[\frac{\beta_1\bar{X}+\beta_2}{\beta_1\bar{x}+\beta_2}\right]$ *Yan and Tian (2010)*	$\theta_{17} = \frac{\beta_1\bar{X}}{\beta_1\bar{X}+\beta_2}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{17}^2 C_x^2 - \theta_{17} \rho C_x C_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{17}^2 C_x^2 - 2\theta_{17} \rho C_x C_y\right)$

Table 1 Continued

Estimator	Constant θ_i	Bias – $B(.)$	Mean squared error MSE(.)
$\hat{\bar{Y}}_{18}=\bar{y}\left[\frac{\beta_2\bar{X}+\beta_1}{\beta_2\bar{x}+\beta_1}\right]$ *Yan and Tian (2010)*	$\theta_{18}=\frac{\beta_2\bar{X}}{\beta_2\bar{X}+\beta_1}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{18}^2C_x^2-\theta_{18}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{18}^2C_x^2-2\theta_{18}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{19}=\bar{y}\left[\frac{\rho\bar{X}+\beta_2}{\rho\bar{x}+\beta_2}\right]$	$\theta_{19}=\frac{\rho\bar{X}}{\rho\bar{X}+\beta_2}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{19}^2C_x^2-\theta_{19}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{19}^2C_x^2-2\theta_{19}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{20}=\bar{y}\left[\frac{\beta_2\bar{X}+\rho}{\beta_2\bar{x}+\rho}\right]$	$\theta_{20}=\frac{\beta_2\bar{X}}{\beta_2\bar{X}+\rho}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{20}^2C_x^2-\theta_{20}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{20}^2C_x^2-2\theta_{20}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{21}=\bar{y}\left[\frac{S_x\bar{X}+\beta_2}{S_x\bar{x}+\beta_2}\right]$	$\theta_{21}=\frac{S_x\bar{X}}{S_x\bar{X}+\beta_2}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{21}^2C_x^2-\theta_{21}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{21}^2C_x^2-2\theta_{21}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{22}=\bar{y}\left[\frac{\beta_2\bar{X}+S_x}{\beta_2\bar{x}+S_x}\right]$ *Singh (2003)*	$\theta_{22}=\frac{\beta_2\bar{X}}{\beta_2\bar{X}+S_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{22}^2C_x^2-\theta_{22}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{22}^2C_x^2-2\theta_{22}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{23}=\bar{y}\left[\frac{M_d\bar{X}+\beta_2}{M_d\bar{x}+\beta_2}\right]$	$\theta_{23}=\frac{M_d\bar{X}}{M_d\bar{X}+\beta_2}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{23}^2C_x^2-\theta_{23}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{23}^2C_x^2-2\theta_{23}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{24}=\bar{y}\left[\frac{\beta_2\bar{X}+M_d}{\beta_2\bar{x}+M_d}\right]$ *Subramani and Kumarapandiyan (2012a)*	$\theta_{24}=\frac{\beta_2\bar{X}}{\beta_2\bar{X}+M_d}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{24}^2C_x^2-\theta_{24}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{24}^2C_x^2-2\theta_{24}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{25}=\bar{y}\left[\frac{\rho\bar{X}+\beta_1}{\rho\bar{x}+\beta_1}\right]$	$\theta_{25}=\frac{\rho\bar{X}}{\rho\bar{X}+\beta_1}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{25}^2C_x^2-\theta_{25}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{25}^2C_x^2-2\theta_{25}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{26}=\bar{y}\left[\frac{\beta_1\bar{X}+\rho}{\beta_1\bar{x}+\rho}\right]$	$\theta_{26}=\frac{\beta_1\bar{X}}{\beta_1\bar{X}+\rho}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{26}^2C_x^2-\theta_{26}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{26}^2C_x^2-2\theta_{26}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{27}=\bar{y}\left[\frac{S_x\bar{X}+\beta_1}{S_x\bar{x}+\beta_1}\right]$	$\theta_{27}=\frac{S_x\bar{X}}{S_x\bar{X}+\beta_1}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{27}^2C_x^2-\theta_{27}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{27}^2C_x^2-2\theta_{27}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{28}=\bar{y}\left[\frac{\beta_1\bar{X}+S_x}{\beta_1\bar{x}+S_x}\right]$ *Singh (2003)*	$\theta_{28}=\frac{\beta_1\bar{X}}{\beta_1\bar{X}+S_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{28}^2C_x^2-\theta_{28}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{28}^2C_x^2-2\theta_{28}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{29}=\bar{y}\left[\frac{M_d\bar{X}+\beta_1}{M_d\bar{x}+\beta_1}\right]$	$\theta_{29}=\frac{M_d\bar{X}}{M_d\bar{X}+\beta_1}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{29}^2C_x^2-\theta_{29}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{29}^2C_x^2-2\theta_{29}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{30}=\bar{y}\left[\frac{\beta_1\bar{X}+M_d}{\beta_1\bar{x}+M_d}\right]$ *Subramani and Kumarapandiyan (2012b)*	$\theta_{30}=\frac{\beta_1\bar{X}}{\beta_1\bar{X}+M_d}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{30}^2C_x^2-\theta_{30}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{30}^2C_x^2-2\theta_{30}\rho C_xC_y\right)$

Table 1 Continued

Estimator	Constant θ_i	Bias – $B(.)$	Mean squared error MSE(.)
$\hat{\bar{Y}}_{31}=\bar{y}\left[\frac{S_x\bar{X}+\rho}{S_x\bar{x}+\rho}\right]$	$\theta_{31}=\frac{S_x\bar{X}}{S_x\bar{X}+\rho}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{31}^2C_x^2-\theta_{31}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{31}^2C_x^2-2\theta_{31}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{32}=\bar{y}\left[\frac{\rho\bar{X}+S_x}{\rho\bar{x}+S_x}\right]$	$\theta_{32}=\frac{\rho\bar{X}}{\rho\bar{X}+S_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{32}^2C_x^2-\theta_{32}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{32}^2C_x^2-2\theta_{32}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{33}=\bar{y}\left[\frac{M_d\bar{X}+\rho}{M_d\bar{x}+\rho}\right]$	$\theta_{33}=\frac{M_d\bar{X}}{M_d\bar{X}+\rho}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{33}^2C_x^2-\theta_{33}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{33}^2C_x^2-2\theta_{33}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{34}=\bar{y}\left[\frac{\rho\bar{X}+M_d}{\rho\bar{x}+M_d}\right]$ *Subramani and Kumarapandiyan (2013b)*	$\theta_{34}=\frac{\rho\bar{X}}{\rho\bar{X}+M_d}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{34}^2C_x^2-\theta_{34}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{34}^2C_x^2-2\theta_{34}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{35}=\bar{y}\left[\frac{M_d\bar{X}+S_x}{M_d\bar{x}+S_x}\right]$	$\theta_{35}=\frac{M_d\bar{X}}{M_d\bar{X}+S_x}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{35}^2C_x^2-\theta_{35}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{35}^2C_x^2-2\theta_{35}\rho C_xC_y\right)$
$\hat{\bar{Y}}_{36}=\bar{y}\left[\frac{S_x\bar{X}+M_d}{S_x\bar{x}+M_d}\right]$	$\theta_{36}=\frac{S_x\bar{X}}{S_x\bar{X}+M_d}$	$\frac{(1-f)}{n}\bar{Y}\left(\theta_{36}^2C_x^2-\theta_{36}\rho C_xC_y\right)$	$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{36}^2C_x^2-2\theta_{36}\rho C_xC_y\right)$

Proposed generalized ratio estimator

As stated earlier, the performance of the estimator of the study variable can be improved by using the known population parameters of the auxiliary variable, which are positively correlated with that of study variable.

The proposed generalized modified ratio estimator for estimating the population mean $\bar{Y}$ is defined as:

$$\hat{\bar{Y}}_{p_i}=\bar{y}\left[\frac{\bar{X}+(1+\alpha)\lambda_i}{\bar{x}+(1+\alpha)\lambda_i}\right];i=1,2,3,\ldots,36 \tag{7}$$

The bias and mean squared error of the proposed estimator $\hat{\bar{Y}}_{p_i}$ have been derived (see Appendix A) and are given below:

$$B\left(\hat{\bar{Y}}_{p_i}\right)=\frac{(1-f)}{n}\bar{Y}\left(\theta_{p_i}^2C_x^2-\theta_{p_i}\rho C_xC_y\right);i=1,2,3,\ldots,36 \tag{8}$$

$$MSE\left(\hat{\bar{Y}}_{p_i}\right)=\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{p_i}^2C_x^2-2\theta_{p_i}\rho C_xC_y\right);$$
$$\text{where } \theta_{p_i}=\frac{\bar{x}}{\bar{x}+(1-\alpha)\lambda_i};i=1,2,3,\ldots,36 \tag{9}$$

where $\lambda_1=C_x,\ \lambda_2=\beta_2,\ \lambda_3=\beta_1,\ \lambda_4=\rho,\ \lambda_5=S_x,\ \lambda_6=M_d,\ \lambda_7=C_x/\beta_2,$ $\lambda_8=\beta_2/C_x,\ \lambda_9=C_x/\beta_1,\ \lambda_{10}=\beta_1/C_x,\ \lambda_{11}=C_x/\rho,\ \lambda_{12}=\rho/C_x,\ \lambda_{13}=C_x/S_x,$ $\lambda_{14}=S_x/C_x,\ \lambda_{15}=C_x/M_d,\ \lambda_{16}=M_d/C_x,\ \lambda_{17}=\beta_2/\beta_1,\ \lambda_{18}=\beta_1/\beta_2,$ $\lambda_{19}=\beta_2/\rho,\ \lambda_{20}=\rho/\beta_2,\ \lambda_{21}=\beta_2/S_x,\ \lambda_{22}=S_x/\beta_2,\ \lambda_{23}=\beta_2/M_d,$ $\lambda_{24}=M_d/\beta_2,\ \lambda_{25}=\beta_1/\rho,\ \lambda_{26}=\rho/\beta_1,\ \lambda_{27}=\beta_1/S_x,\ \lambda_{28}=S_x/\beta_1,$ $\lambda_{29}=\beta_1/M_d,\ \lambda_{30}=M_d/\beta_1,\ \lambda_{31}=\rho/S_x,\ \lambda_{32}=S_x/\rho,\ \lambda_{33}=\rho/M_d,$ $\lambda_{34}=M_d/\rho,\ \lambda_{35}=S_x/M_d,$ and $\lambda_{36}=M_d/S_x$

Efficiency of the proposed estimator

The variance of SRSWOR sample mean $\bar{y}_{srs}$ is given below:

$$V\left(\bar{y}_{srs}\right)=\frac{(1-f)}{n}S_y^2 \tag{10}$$

The bias and mean squared error of the usual ratio estimator $\hat{\bar{Y}}_R$ to the first degree of approximation are given below:

$$B\left(\hat{\bar{Y}}_R\right)=\frac{(1-f)}{n}\bar{Y}\left(C_x^2-\rho C_xC_y\right)$$
$$MSE\left(\hat{\bar{Y}}_R\right)=\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+C_x^2-2\rho C_xC_y\right) \tag{11}$$

The bias and the mean squared error of the modified ratio estimators $\hat{\bar{Y}}_1$ to $\hat{\bar{Y}}_{36}$ listed in the Table 1 are represented in a single class as given below:

$$\hat{\bar{Y}}_i = \bar{y}\left[\frac{\bar{X}+\lambda_i}{\bar{x}+\lambda_i}\right]; i = 1,2,3,\ldots,36$$
$$B\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)}{n}\bar{Y}\left(\theta_i^2 C_x^2 - \rho\theta_i C_x C_y\right); i = 1,2,3,\ldots,36 \quad (12)$$
$$MSE\left(\hat{\bar{Y}}_i\right) = \frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_i^2 C_x^2 - 2\rho\theta_i C_x C_y\right) \text{ where } \theta_i = \frac{\bar{X}}{\bar{X}+\lambda_i}; i = 1,2,3,\ldots,36$$

As discussed earlier, the bias, the mean squared error and the constant of the proposed modified ratio estimator $\hat{\bar{Y}}_{p_i}$ are given below:

$$B\left(\hat{\bar{Y}}_{p_i}\right) = \frac{(1-f)}{n}\bar{Y}\left(\theta_{p_i}^2 C_x^2 - \rho\theta_{p_i} C_x C_y\right); i = 1,2,3,\ldots,36$$
$$MSE\left(\hat{\bar{Y}}_{p_i}\right) = \frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{p_i}^2 C_x^2 - 2\rho\theta_{p_i} C_x C_y\right) \quad (13)$$
$$\text{where } \theta_{p_i} = \frac{\bar{X}}{\bar{X}+(1+\alpha)\lambda_i}; i = 1,2,3,\ldots,36$$

where $\lambda_1 = C_x$, $\lambda_2 = \beta_2$, $\lambda_3 = \beta_1$, $\lambda_4 = \rho$, $\lambda_5 = S_x$, $\lambda_6 = M_d$, $\lambda_7 = C_x/\beta_2$, $\lambda_8 = \beta_2/C_x$, $\lambda_9 = C_x/\beta_1$, $\lambda_{10} = \beta_1/C_x$, $\lambda_{11} = C_x/\rho$, $\lambda_{12} = \rho/C_x$, $\lambda_{13} = C_x/S_x$, $\lambda_{14} = S_x/C_x$, $\lambda_{15} = C_x/M_d$, $\lambda_{16} = M_d/C_x$, $\lambda_{17} = \beta_2/\beta_1$, $\lambda_{18} = \beta_1/\beta_2$, $\lambda_{19} = \beta_2/\rho$, $\lambda_{20} = \rho/\beta_2$, $\lambda_{21} = \beta_2/S_x$, $\lambda_{22} = S_x/\beta_2$, $\lambda_{23} = \beta_2/M_d$, $\lambda_{24} = M_d/\beta_2$, $\lambda_{25} = \beta_1/\rho$, $\lambda_{26} = \rho/\beta_1$, $\lambda_{27} = \beta_1/S_x$, $\lambda_{28} = S_x/\beta_1$, $\lambda_{29} = \beta_1/M_d$, $\lambda_{30} = M_d/\beta_1$, $\lambda_{31} = \rho/S_x$, $\lambda_{32} = S_x/\rho$, $\lambda_{33} = \rho/M_d$, $\lambda_{34} = M_d/\rho$, $\lambda_{35} = S_x/M_d$, and $\lambda_{36} = M_d/S_x$

From the expressions given in (10) and (13), the conditions (see Appendix B) for which the proposed estimator $\hat{\bar{Y}}_{p_i}$ are more efficient than the simple random sampling without replacement (SRSWOR) sample mean $\bar{y}_{srs}$ were derived and are:

$$MSE\left(\hat{\bar{Y}}_{p_i}\right) \le V\left(\bar{y}_r\right) \text{ if } \theta_{p_i} \le 2\rho\frac{C_y}{C_x} \quad (14)$$

From the expressions given in (11) and (13), the conditions (see Appendix C) for which the proposed estimators $\hat{\bar{Y}}_{p_i}$ are more efficient than the usual ratio estimator $\hat{\bar{Y}}_R$ were derived and are:

$$MSE\left(\hat{\bar{Y}}_{p_i}\right) \le MSE\left(\hat{\bar{Y}}_R\right) \text{ either } \frac{2\rho C_y}{C_x} - 1 \le \theta_{p_i} \le 1 \text{ (or) } 1 \le \theta_{p_i} \le \frac{2\rho C_y}{C_x} - 1 \quad (15)$$

From the expressions given in (12) and (13), the conditions (see Appendix D) for which the proposed estimators $\hat{\bar{Y}}_{p_j}; j = 1, 2, \ldots, 5$ are more efficient than the existing modified ratio estimators given in Class 1, $\bar{Y}_i; i = 1, 2, 3, \ldots, 11$ were derived and are:

$$MSE\left(\hat{\bar{Y}}_{p_j}\right) \le MSE\left(\hat{\bar{Y}}_i\right) \text{ either } \frac{2\rho C_y}{C_x} - \theta_i \le \theta_{p_j} \le \theta_i \text{ (or) } \theta_i \le \theta_{p_j} \le \frac{2\rho C_y}{C_x} - \theta_i \quad (16)$$

The conditions in terms of α in which proposed estimator $\hat{\bar{Y}}_{p_i}$ performs better than the simple random sampling without replacement (SRSWOR) sample mean $\bar{y}_{srs}$ were obtained and are:

$$MSE\left(\hat{\bar{Y}}_{p_i}\right) \le V\left(\bar{y}_r\right) \text{ if } \alpha_i \ge \frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right] \quad (17)$$

From the expression given in (15), the range of α in which proposed estimator $\hat{\bar{Y}}_{p_i}$ performs better than the usual ratio estimator $\hat{\bar{Y}}_R$ is determined and is:

$$MSE\left(\hat{\bar{Y}}_{p_i}\right) \le MSE\left(\hat{\bar{Y}}_R\right) \text{ either } -1 \le \alpha_i \le \frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho - 1\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right]$$

$$\text{(or)} \quad (18)$$

$$\frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho - 1\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right] \le \alpha_i \le -1; i = 1, 2, 3, \ldots, 36$$

From the expression given in (16), the range of α in which proposed estimator $\hat{\bar{Y}}_{p_i}$ performs better than the existing modified ratio estimators listed in Table 1 is:

$$MSE\left(\hat{\bar{Y}}_{p_i}\right) \le MSE\left(\hat{\bar{Y}}_i\right) \text{ either } 0 \le \alpha_i \le \frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho - \theta_i\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right]$$

$$\text{(or)} \qquad (19)$$

$$\frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho - \theta_i\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right] \le \alpha_i \le 0; i = 1,2,3,\ldots,36$$

Particular case:

1) At $\alpha_i = \frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right]; i = 1,2,3,\ldots,36,$ the mean squared error of the proposed estimator $\hat{\bar{Y}}_{p_i}; i = 1,2,3,\ldots,36$ equal to the variance of the SRSWOR sample mean $\bar{y}_{srs}$.

2) At limit point $\alpha_i = \frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho - 1\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right]$ or -1 the mean squared error of the proposed estimator $\hat{\bar{Y}}_{p_i}; i = 1,2,3,\ldots,36$ equal to the mean squared error of the usual ratio estimator $\hat{\bar{Y}}_R$

3) At limit point $\frac{\bar{X}}{\lambda_i}\left[\left(2\frac{C_y}{C_x}\rho - \theta_i\right)^{-1} - 1 - \frac{\lambda_i}{\bar{X}}\right]$ or 0 the mean squared error of the proposed estimator $\hat{\bar{Y}}_{p_i}; i = 1,2,3,\ldots,36$ the mean squared error of the existing modified ratio estimators $\hat{\bar{Y}}_i; i = 1,2,3,\ldots,36$

4) At $\alpha_i = \frac{\bar{X}}{\lambda_i}\left[\left(\frac{C_y}{C_x}\rho\right)^{-1} - 1\right] - 1; i = 1,2,3,\ldots,36,$ the means squared error of the proposed estimator $\hat{\bar{Y}}_{p_i}; i = 1,2,3,\ldots,36$ equal to the variance of the usual linear regression estimator $\hat{\bar{Y}}_{1r}$

Numerical Study

The performance of the proposed generalized modified ratio estimator is assessed with that of the SRSWOR sample mean, the usual ratio estimator and the existing modified ratio estimators listed in Table 1 for certain natural populations. In this connection, four natural populations for the assessment of the performance of the proposed estimators with that of existing estimators were considered. Population 1 is taken from Singh and Chaudhary (1986) given in page 108; population 2 and population 3 are taken from Singh and Chaudhary (1986) given in page 177; population 4 is taken from Cochran (1977) given in page 152. The population parameters and the constants computed from the above populations are given below in Table 2, whereas the range of α in which proposed estimator performs better than the existing estimators, the constants, the biases and the mean squared errors of the existing and proposed estimators for the above populations are respectively given from the Tables 3 to 8.

Table 2. Parameters and constants of the population

Parameters	Population 1	Population 2	Population 3	Population 4
N	70	34	34	49
n	25	20	20	20
$\overline{Y}$	96.7000	856.4118	85.6412	127.7959
$\overline{X}$	175.2671	208.8824	19.9441	103.1429
ρ	0.7293	0.4491	0.4453	0.9817
S_y	60.4714	733.1407	73.3141	123.1212
C_y	0.6254	0.8561	0.8561	0.9634
S_x	140.8572	150.5060	15.0215	104.4051
C_x	0.8037	0.7205	0.7532	1.0122
$\beta_2(x)$	7.0952	0.0974	3.7257	7.5114
$\beta_1(x)$	1.9507	0.9782	1.1823	2.2553
M_d	121.5000	150.0000	14.2500	64.0000

Table 3. Range of α in which proposed estimator performs better than the usual ratio estimator

Estimator	α range (α_L, α_u)			
	Population 1	**Population 2**	**Population 3**	**Population 4**
$\hat{\bar{Y}}_{p_1}$	(-1, 1396.1641)	(-1, 4023.1475)	(-1, 2138.4916)	(-1, 14.3885)
$\hat{\bar{Y}}_{p_2}$	(-1, 157.2620)	(-1, 29751.6396)	(-1, 431.5268)	(-1, 1.0738)
$\hat{\bar{Y}}_{p_3}$	(-1, 574.6399)	(-1, 2963.2549)	(-1, 1361.9917)	(-1, 5.9068)
$\hat{\bar{Y}}_{p_4}$	(-1, 1538.6966)	(-1, 6454.9931)	(-1, 3617.8302)	(-1, 14.8665)
$\hat{\bar{Y}}_{p_5}$	(-1, 6.9719)	(-1, 18.2651)	(-1, 106.2772)	(-1, -0.8508)
$\hat{\bar{Y}}_{p_6}$	(-1, 8.2420)	(-1, 18.3301)	(-1, 112.0853)	(-1, -0.7566)
$\hat{\bar{Y}}_{p_7}$	(-1, 9912.1584)	(-1, 391.1699)	(-1, 7970.1040)	(-1, 114.5893)
$\hat{\bar{Y}}_{p_8}$	(-1, 126.1952)	(-1, 21436.6645)	(-1, 324.7792)	(-1, 1.0991)
$\hat{\bar{Y}}_{p_9}$	(-1, 2724.4479)	(-1, 3935.2639)	(-1, 2528.5210)	(-1, 33.7053)
$\hat{\bar{Y}}_{p_{10}}$	(-1, 461.6418)	(-1, 2134.8341)	(-1, 1025.6054)	(-1, 5.9914)
$\hat{\bar{Y}}_{p_{11}}$	(-1, 1017.9517)	(-1, 1806.3268)	(-1, 951.7156)	(-1, 14.1076)
$\hat{\bar{Y}}_{p_{12}}$	(-1, 1236.4542)	(-1, 4650.7357)	(-1, 2724.7029)	(-1, 15.0607)
$\hat{\bar{Y}}_{p_{13}}$	(-1, 196799.6166)	(-1, 605657.2149)	(-1, 32137.3735)	(-1, 1605.6390)
$\hat{\bar{Y}}_{p_{14}}$	(-1, 5.4070)	(-1, 12.8811)	(-1, 79.8012)	(-1, -0.8490)
$\hat{\bar{Y}}_{p_{15}}$	(-1, 169754.4325)	(-1, 603621.1259)	(-1, 30486.7557)	(-1, 983.8649)
$\hat{\bar{Y}}_{p_{16}}$	(-1, 6.4278)	(-1, 12.9279)	(-1, 84.1758)	(-1, -0.7536)
$\hat{\bar{Y}}_{p_{17}}$	(-1, 307.7217)	(-1, 29101.8696)	(-1, 510.3764)	(-1, 3.6769)
$\hat{\bar{Y}}_{p_{18}}$	(-1, 4083.2802)	(-1, 287.8790)	(-1, 5077.0982)	(-1, 50.8800)
$\hat{\bar{Y}}_{p_{19}}$	(-1, 114.4205)	(-1, 13361.518)	(-1, 191.6042)	(-1, 1.0359)
$\hat{\bar{Y}}_{p_{20}}$	(-1, 10923.4555)	(-1, 628.1634)	(-1, 13481.6757)	(-1, 118.1797)
$\hat{\bar{Y}}_{p_{21}}$	(-1, 22291.3463)	(-1, 4477948.8100)	(-1, 6496.2013)	(-1, 215.5109)

Table 3 continued.

Estimator	α range (α_L, α_u)			
	Population 1	**Population 2**	**Population 3**	**Population 4**
$\hat{\bar{Y}}_{p_{22}}$	(-1, 55.5623)	(-1, 0.8775)	(-1, 398.6828)	(-1, 0.1207)
$\hat{\bar{Y}}_{p_{23}}$	(-1, 19227.8366)	(-1, 4462894.9330)	(-1, 6162.5069)	(-1, 131.7205)
$\hat{\bar{Y}}_{p_{24}}$	(-1, 64.5737)	(-1, 0.8838)	(-1, 420.3218)	(-1, 0.8282)
$\hat{\bar{Y}}_{p_{25}}$	(-1, 418.8142)	(-1, 1330.3074)	(-1, 605.9402)	(-1, 5.7807)
$\hat{\bar{Y}}_{p_{26}}$	(-1, 3002.4862)	(-1, 6314.0002)	(-1, 4277.5430)	(-1, 34.7834)
$\hat{\bar{Y}}_{p_{27}}$	81082.0243)	(-1, 446137.0551)	(-1, 20473.1799)	(-1, 720.1090)
$\hat{\bar{Y}}_{p_{28}}$	(-1, 14.5508)	(-1, 17.8444)	(-1, 125.8339)	(-1, -0.6635)
$\hat{\bar{Y}}_{p_{29}}$	(-1, 69939.2477)	(-1, 444637.2376)	(-1, 19421.6318)	(-1, 441.0376)
$\hat{\bar{Y}}_{p_{30}}$	(-1, 17.0283)	(-1, 17.908)	(-1, 132.7007)	(-1, -0.4511)
$\hat{\bar{Y}}_{p_{31}}$	(-1, 216876.3557)	(-1, 971664.4899)	(-1, 54359.2581)	(-1, 1655.5450)
$\hat{\bar{Y}}_{p_{32}}$	(-1, 4.8139)	(-1, 7.6524)	(-1, 46.7706)	(-1, -0.8535)
$\hat{\bar{Y}}_{p_{33}}$	(-1, 187072.1402)	(-1, 968397.9653)	(-1, 51567.3306)	(-1, 1014.4570)
$\hat{\bar{Y}}_{p_{34}}$	(-1, 5.7402)	(-1, 7.6816)	(-1, 49.3569)	(-1, -0.7611)
$\hat{\bar{Y}}_{p_{35}}$	(-1, 967.5869)	(-1, 2888.7708)	(-1, 1527.7007)	(-1, 8.5485)
$\hat{\bar{Y}}_{p_{36}}$	(-1, 1300.7996)	(-1, 2908.2988)	(-1, 1697.7104)	(-1, 24.4109)

Table 4. Range of α in which proposed estimator performs better than the existing modified ratio estimators

Estimator	α range (α_L, α_u)			
	Population 1	**Population 2**	**Population 3**	**Population 4**
$\hat{\bar{Y}}_{p_1}$ w.r.t. $\hat{\bar{Y}}_1$	(0, 1343.3398)	(0, 3813.2012)	(0, 517.1768)	(0, 13.0911)
$\hat{\bar{Y}}_{p_2}$ w.r.t. $\hat{\bar{Y}}_2$	(0, 116.3312)	(0, 29531.8209)	(0, 25.2048)	(-0.0717, 0)
$\hat{\bar{Y}}_{p_3}$ w.r.t. $\hat{\bar{Y}}_3$	(0, 524.4732)	(0, 2757.1363)	(0, 229.5184)	(0, 4.6415)
$\hat{\bar{Y}}_{p_4}$ w.r.t. $\hat{\bar{Y}}_4$	(0, 1485.6902)	(0, 6240.8577)	(0, 1268.9984)	(0, 13.5682)
$\hat{\bar{Y}}_{p_5}$ w.r.t. $\hat{\bar{Y}}_5$	(-0.1011, 0)	(0, 0.4679)	(0, 0.6773)	(-1.2678, 0)
$\hat{\bar{Y}}_{p_6}$ w.r.t. $\hat{\bar{Y}}_6$	(0, 0.2071)	(0, 0.4777)	(0, 0.8631)	(-1.3241, 0)
$\hat{\bar{Y}}_{p_7}$ w.r.t. $\hat{\bar{Y}}_7$	(0, 9857.5942)	(0, 249.4093)	(0, 4332.2435)	(0, 113.2681)
$\hat{\bar{Y}}_{p_8}$ w.r.t. $\hat{\bar{Y}}_8$	(0, 87.6545)	(0, 21217.4712)	(0, 14.0810)	(-0.0482, 0)
$\hat{\bar{Y}}_{p_9}$ w.r.t. $\hat{\bar{Y}}_9$	(0, 2670.6504)	(0, 3725.5608)	(0, 692.7328)	(0, 32.3928)
$\hat{\bar{Y}}_{p_{10}}$ w.r.t. $\hat{\bar{Y}}_{10}$	(0, 412.5020)	(0, 1934.1048)	(0, 135.4434)	(0, 4.7253)
$\hat{\bar{Y}}_{p_{11}}$ w.r.t. $\hat{\bar{Y}}_{11}$	(0, 965.8454)	(0, 1608.9539)	(0, 117.6363)	(0, 12.8106)
$\hat{\bar{Y}}_{p_{12}}$ w.r.t. $\hat{\bar{Y}}_{12}$	(0, 1183.8818)	(0, 4439.3079)	(0, 787.7525)	(0, 13.7621)
$\hat{\bar{Y}}_{p_{13}}$ w.r.t. $\hat{\bar{Y}}_{13}$	(0, 196744.7709)	(0, 605435.8482)	(0, 26599.1484)	(0, 1604.3148)
$\hat{\bar{Y}}_{p_{14}}$ w.r.t. $\hat{\bar{Y}}_{14}$	(-0.4252, 0)	(-0.2370, 0)	(-0.0477, 0)	(-1.2694, 0)
$\hat{\bar{Y}}_{p_{15}}$ w.r.t. $\hat{\bar{Y}}_{15}$	(0, 169699.5893)	(0, 603399.7594)	(0, 24999.7154)	(0, 982.5405)
$\hat{\bar{Y}}_{p_{16}}$ w.r.t. $\hat{\bar{Y}}_{16}$	(-0.2210, 0)	(-0.2317, 0)	(0, 0.0582)	(-1.3250, 0)
$\hat{\bar{Y}}_{p_{17}}$ w.r.t. $\hat{\bar{Y}}_{17}$	(0, 261.0105)	(0, 28882.0870)	(0, 35.2418)	(0, 2.4381)
$\hat{\bar{Y}}_{p_{18}}$ w.r.t. $\hat{\bar{Y}}_{18}$	(0, 4029.1336)	(0, 162.2788)	(0, 2189.8216)	(0, 49.5634)
$\hat{\bar{Y}}_{p_{19}}$ w.r.t. $\hat{\bar{Y}}_{19}$	(0, 77.0150)	(0, 13143.6647)	(0, 4.3618)	(-0.1066, 0)
$\hat{\bar{Y}}_{p_{20}}$ w.r.t. $\hat{\bar{Y}}_{20}$	(0, 10868.8640)	(0, 464.1746)	(0, 9009.6165)	(0, 116.8585)
$\hat{\bar{Y}}_{p_{21}}$ w.r.t. $\hat{\bar{Y}}_{21}$	(0, 22236.6179)	(0, 4477727.3738)	(0, 3199.5909)	(0, 214.1880)

Table 4 continued.

Estimator	α range (α_L, α_u)			
	Population 1	**Population 2**	**Population 3**	**Population 4**
$\hat{\bar{Y}}_{p_{22}}$ w.r.t. $\hat{\bar{Y}}_{22}$	(0, 27.4608)	(-0.9926, 0)	(0, 21.4756)	(-0.9064, 0)
$\hat{\bar{Y}}_{p_{23}}$ w.r.t. $\hat{\bar{Y}}_{23}$	(0, 19173.1292)	(0, 4462673.4962)	(0, 2954.0212)	(0, 130.3989)
$\hat{\bar{Y}}_{p_{24}}$ w.r.t. $\hat{\bar{Y}}_{24}$	(0, 34.4533)	(-0.9925, 0)	(0, 23.9023)	(-0.2968, 0)
$\hat{\bar{Y}}_{p_{25}}$ w.r.t. $\hat{\bar{Y}}_{25}$	(0, 370.1916)	(0, 1140.3196)	(0, 49.3999)	(0, 4.5165)
$\hat{\bar{Y}}_{p_{26}}$ w.r.t. $\hat{\bar{Y}}_{26}$	(0, 2948.5920)	(0, 6100.0225)	(0, 1667.5443)	(0, 33.4704)
$\hat{\bar{Y}}_{p_{27}}$ w.r.t. $\hat{\bar{Y}}_{27}$	(0, 81027.2001)	(0, 445915.7171)	(0, 15429.9892)	(0, 718.7848)
$\hat{\bar{Y}}_{p_{28}}$ w.r.t. $\hat{\bar{Y}}_{28}$	(0, 2.2599)	(0, 0.4053)	(0, 1.3412)	(-1.3376, 0)
$\hat{\bar{Y}}_{p_{29}}$ w.r.t. $\hat{\bar{Y}}_{29}$	(0, 69884.4292)	(0, 444415.8999)	(0, 14444.8014)	(0, 439.7137)
$\hat{\bar{Y}}_{p_{30}}$ w.r.t. $\hat{\bar{Y}}_{30}$	(0, 3.2704)	(0, 0.4146)	(0, 1.6000)	(-1.2834, 0)
$\hat{\bar{Y}}_{p_{31}}$ w.r.t. $\hat{\bar{Y}}_{31}$	(0, 216821.5087)	(0, 971443.0928)	(0, 48401.3982)	(0, 1654.2204)
$\hat{\bar{Y}}_{p_{32}}$ w.r.t. $\hat{\bar{Y}}_{32}$	(-0.5310, 0)	(-0.7110, 0)	(-0.6684, 0)	(-1.2652, 0)
$\hat{\bar{Y}}_{p_{33}}$ w.r.t. $\hat{\bar{Y}}_{33}$	(0, 187017.2953)	(0, 968176.5684)	(0, 45644.6094)	(0, 1013.1325)
$\hat{\bar{Y}}_{p_{34}}$ w.r.t. $\hat{\bar{Y}}_{34}$	(-0.3616, 0)	(-0.7090, 0)	(-0.6313, 0)	(-1.3225, 0)
$\hat{\bar{Y}}_{p_{35}}$ w.r.t. $\hat{\bar{Y}}_{35}$	(0, 915.6162)	(0, 2683.0193)	(0, 283.1533)	(0, 7.2673)
$\hat{\bar{Y}}_{p_{36}}$ w.r.t. $\hat{\bar{Y}}_{36}$	(0, 1248.1186)	(0, 2702.4493)	(0, 342.7819)	(0, 23.1028)

Table 5. Constant, Bias and Mean squared error of the Existing and Proposed estimators for Population 1

Estimator	θ_i	$B\left(\hat{\bar{Y}}_{(.)}\right)$	$MSE\left(\hat{\bar{Y}}_{(.)}\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\bar{y}_{srs}$	-	-	94.0466	94.0466	-	-	-
$\hat{\bar{Y}}_R$	-	0.6946	73.0773	73.0773	-	-	-
$\hat{\bar{Y}}_{p_1}$	0.9954	0.6842	72.4673	72.4673	0.2448	0.1269	60.1973
$\hat{\bar{Y}}_{p_2}$	0.9611	0.6076	68.0853	68.0853	0.2945	0.1291	55.5981
$\hat{\bar{Y}}_{p_3}$	0.9890	0.6695	71.6173	71.6173	0.2545	0.1279	59.2452
$\hat{\bar{Y}}_{p_4}$	0.9959	0.6851	72.5232	72.5232	0.2442	0.1268	60.2610
$\hat{\bar{Y}}_{p_5}$	0.5544	0.0116	44.0518	44.0518	0.5672	0.0003	44.0253
$\hat{\bar{Y}}_{p_6}$	0.5906	0.0219	44.1080	44.1080	0.5666	0.0009	44.0254
$\hat{\bar{Y}}_{p_7}$	0.9994	0.6932	72.9906	72.9906	0.2389	0.1261	60.7979
$\hat{\bar{Y}}_{p_8}$	0.9520	0.5880	66.9919	66.9919	0.3069	0.1285	54.5701
$\hat{\bar{Y}}_{p_9}$	0.9977	0.6893	72.7631	72.7631	0.2415	0.1264	60.5354
$\hat{\bar{Y}}_{p_{10}}$	0.9863	0.6635	71.2712	71.2712	0.2584	0.1283	58.8657
$\hat{\bar{Y}}_{p_{11}}$	0.9938	0.6803	72.2439	72.2439	0.2474	0.1272	59.9443
$\hat{\bar{Y}}_{p_{12}}$	0.9948	0.6828	72.3894	72.3894	0.2457	0.1270	60.1089
$\hat{\bar{Y}}_{p_{13}}$	1.0000	0.6946	73.0729	73.0729	0.2380	0.1260	60.8934
$\hat{\bar{Y}}_{p_{14}}$	0.5000	0.0542	44.7328	44.7328	0.5595	0.0072	44.0353
$\hat{\bar{Y}}_{p_{15}}$	1.0000	0.6946	73.0722	73.0722	0.2380	0.1260	60.8925
$\hat{\bar{Y}}_{p_{16}}$	0.5369	0.0264	44.1707	44.1707	0.5659	0.0015	44.0257
$\hat{\bar{Y}}_{p_{17}}$	0.9797	0.6485	70.4101	70.4101	0.2682	0.1289	57.9425
$\hat{\bar{Y}}_{p_{18}}$	0.9984	0.6911	72.8672	72.8672	0.2403	0.1263	60.6553
$\hat{\bar{Y}}_{p_{19}}$	0.9474	0.5781	66.4416	66.4416	0.3132	0.1279	54.0709

Table 5 continued.

Estimator	θ_i	$B\left(\hat{\bar{Y}}_{(.)}\right)$	$MSE\left(\hat{\bar{Y}}_{(.)}\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\hat{\bar{Y}}_{p_{20}}$	0.9994	0.6933	72.9986	72.9986	0.2388	0.1261	60.8072
$\hat{\bar{Y}}_{p_{21}}$	0.9997	0.6940	73.0387	73.0387	0.2383	0.1260	60.8536
$\hat{\bar{Y}}_{p_{22}}$	0.8983	0.4772	61.0160	61.0160	0.3747	0.1160	49.7965
$\hat{\bar{Y}}_{p_{23}}$	0.9997	0.6939	73.0325	73.0325	0.2384	0.1260	60.8465
$\hat{\bar{Y}}_{p_{24}}$	0.9110	0.5026	62.3499	62.3499	0.3596	0.1201	50.7382
$\hat{\bar{Y}}_{p_{25}}$	0.9850	0.6604	71.0929	71.0929	0.2604	0.1284	58.6722
$\hat{\bar{Y}}_{p_{26}}$	0.9979	0.6898	72.7920	72.7920	0.2411	0.1264	60.5687
$\hat{\bar{Y}}_{p_{27}}$	0.9999	0.6945	73.0667	73.0667	0.2380	0.1260	60.8861
$\hat{\bar{Y}}_{p_{28}}$	0.7082	0.1601	47.1006	47.1006	0.5326	0.0298	44.2143
$\hat{\bar{Y}}_{p_{29}}$	0.9999	0.6944	73.0650	73.0650	0.2380	0.1260	60.8841
$\hat{\bar{Y}}_{p_{30}}$	0.7378	0.2018	48.5296	48.5296	0.5164	0.0424	44.4309
$\hat{\bar{Y}}_{p_{31}}$	1.0000	0.6946	73.0733	73.0733	0.2379	0.1260	60.8938
$\hat{\bar{Y}}_{p_{32}}$	0.4757	-0.0701	45.3331	45.3331	0.5527	0.0132	44.0594
$\hat{\bar{Y}}_{p_{33}}$	1.0000	0.6946	73.0727	73.0727	0.2380	0.1260	60.8931
$\hat{\bar{Y}}_{p_{34}}$	0.5127	0.0451	44.4921	44.4921	0.5622	0.0048	44.0296
$\hat{\bar{Y}}_{p_{35}}$	0.9934	0.6796	72.2012	72.2012	0.2478	0.1272	59.8961
$\hat{\bar{Y}}_{p_{36}}$	0.9951	0.6834	72.4230	72.4230	0.2453	0.1269	60.1471

Table 6. Constant, Bias and Mean squared error of the Existing and Proposed estimators for Population 2

Estimator	θ_i	$B\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\bar{y}_{srs}$	-	-	11066.0800	11066.0800	-	-	-
$\hat{\bar{Y}}_R$	-	4.2694	10539.2700	10539.2700	-	-	-
$\hat{\bar{Y}}_{p_1}$	0.9966	4.2233	10514.2250	10514.2250	0.1319	0.4851	10098.8070
$\hat{\bar{Y}}_{p_2}$	0.9995	4.2631	10535.8620	10535.8620	0.1268	0.4721	10131.5920
$\hat{\bar{Y}}_{p_3}$	0.9953	4.2070	10505.3560	10505.3560	0.1340	0.4903	10085.4900
$\hat{\bar{Y}}_{p_4}$	0.9979	4.2406	10523.6170	10523.6170	0.1297	0.4795	10112.9870
$\hat{\bar{Y}}_{p_5}$	0.5812	0.2533	8851.7250	8851.7250	0.5294	0.0206	8834.0910
$\hat{\bar{Y}}_{p_6}$	0.5820	0.2581	8852.3420	8852.3420	0.5292	0.0213	8834.1010
$\hat{\bar{Y}}_{p_7}$	0.9658	3.8212	10298.4430	10298.4430	0.1835	0.5881	9794.7990
$\hat{\bar{Y}}_{p_8}$	0.9994	4.2607	10534.5420	10534.5420	0.1271	0.4729	10129.5790
$\hat{\bar{Y}}_{p_9}$	0.9965	4.2223	10513.6700	10513.6700	0.1321	0.4854	10097.9710
$\hat{\bar{Y}}_{p_{10}}$	0.9935	4.1831	10492.3780	10492.3780	0.1371	0.4977	10066.1290
$\hat{\bar{Y}}_{p_{11}}$	0.9924	4.1676	10483.9890	10483.9890	0.1392	0.5024	10053.6950
$\hat{\bar{Y}}_{p_{12}}$	0.9970	4.2295	10517.5830	10517.5830	0.1311	0.4831	10103.8680
$\hat{\bar{Y}}_{p_{13}}$	1.0000	4.2691	10539.1030	10539.1030	0.1260	0.4701	10136.5390
$\hat{\bar{Y}}_{p_{14}}$	0.5000	0.1538	8842.8000	8842.8000	0.5315	0.0103	8833.9850
$\hat{\bar{Y}}_{p_{15}}$	1.0000	4.2691	10539.1020	10539.1020	0.1260	0.4701	10136.5380
$\hat{\bar{Y}}_{p_{16}}$	0.5008	0.1502	8842.3620	8842.3620	0.5316	0.0098	8833.9810
$\hat{\bar{Y}}_{p_{17}}$	0.9995	4.2630	10535.7860	10535.7860	0.1268	0.4721	10131.4760
$\hat{\bar{Y}}_{p_{18}}$	0.9542	3.6732	10220.4740	10220.4740	0.2021	0.6133	9695.2120
$\hat{\bar{Y}}_{p_{19}}$	0.9990	4.2555	10531.6900	10531.6900	0.1277	0.4746	10125.2380

Table 6 continued

Estimator	θ_i	$B\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\hat{\bar{Y}}_{p_{20}}$	0.9784	3.9839	10385.0680	10385.0680	0.1628	0.5526	9911.8290
$\hat{\bar{Y}}_{p_{21}}$	1.0000	4.2693	10539.2480	10539.2480	0.1259	0.4700	10136.7600
$\hat{\bar{Y}}_{p_{22}}$	0.1191	0.4520	10180.5970	10180.5970	0.2117	0.6238	9646.3850
$\hat{\bar{Y}}_{p_{23}}$	1.0000	4.2693	10539.2480	10539.2480	0.1259	0.4700	10136.7600
$\hat{\bar{Y}}_{p_{24}}$	0.1195	0.4530	10178.2990	10178.2990	0.2122	0.6243	9643.6140
$\hat{\bar{Y}}_{p_{25}}$	0.9897	4.1318	10464.6450	10464.6450	0.1438	0.5130	10025.2640
$\hat{\bar{Y}}_{p_{26}}$	0.9978	4.2400	10523.2690	10523.2690	0.1298	0.4797	10112.4600
$\hat{\bar{Y}}_{p_{27}}$	1.0000	4.2690	10539.0430	10539.0430	0.1260	0.4701	10136.4470
$\hat{\bar{Y}}_{p_{28}}$	0.5758	0.2226	8847.9320	8847.9320	0.5303	0.0162	8834.0370
$\hat{\bar{Y}}_{p_{29}}$	1.0000	4.2690	10539.0420	10539.0420	0.1260	0.4701	10136.4460
$\hat{\bar{Y}}_{p_{30}}$	0.5767	0.2273	8848.4820	8848.4820	0.5301	0.0169	8834.0440
$\hat{\bar{Y}}_{p_{31}}$	1.0000	4.2692	10539.1660	10539.1660	0.1260	0.4700	10136.6350
$\hat{\bar{Y}}_{p_{32}}$	0.3840	0.5259	9009.4490	9009.4490	0.4916	0.1888	8847.7480
$\hat{\bar{Y}}_{p_{33}}$	1.0000	4.2692	10539.1660	10539.1660	0.1260	0.4700	10136.6350
$\hat{\bar{Y}}_{p_{34}}$	0.3848	0.5242	9007.5850	9007.5850	0.4921	0.1870	8847.4570
$\hat{\bar{Y}}_{p_{35}}$	0.9952	4.2054	10504.4910	10504.4910	0.1343	0.4908	10084.1940
$\hat{\bar{Y}}_{p_{36}}$	0.9953	4.2058	10504.7220	10504.7220	0.1342	0.4906	10084.5400

Table 7. Constant, Bias and Mean squared error of the Existing and Proposed estimators for Population 3

Estimator	θ_i	$B\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\bar{y}_{srs}$	-	-	379.4085	379.4085	-	-	-
$\hat{\bar{Y}}_R$	-	1.6938	375.8179	375.8179	-	-	-
$\hat{\bar{Y}}_{p_1}$	0.9636	1.5119	365.6490	365.6490	0.0926	0.1313	354.4061
$\hat{\bar{Y}}_{p_2}$	0.8426	0.9723	337.4291	337.4291	0.2824	0.2167	318.8736
$\hat{\bar{Y}}_{p_3}$	0.9440	1.4178	360.5017	360.5017	0.1272	0.1653	346.3464
$\hat{\bar{Y}}_{p_4}$	0.9782	1.5835	369.6218	369.6218	0.0658	0.0994	361.1082
$\hat{\bar{Y}}_{p_5}$	0.5704	0.1257	305.3883	305.3883	0.4979	0.0139	304.1944
$\hat{\bar{Y}}_{p_6}$	0.5833	0.1543	305.9229	305.9229	0.4944	0.0199	304.2154
$\hat{\bar{Y}}_{p_7}$	0.9900	1.6427	372.9362	372.9362	0.0435	0.0691	367.0206
$\hat{\bar{Y}}_{p_8}$	0.8013	0.8111	329.7622	329.7622	0.3340	0.1972	312.8772
$\hat{\bar{Y}}_{p_9}$	0.9690	1.5385	367.1190	367.1190	0.0827	0.1201	356.8370
$\hat{\bar{Y}}_{p_{10}}$	0.9270	1.3383	356.2136	356.2136	0.1560	0.1873	340.1698
$\hat{\bar{Y}}_{p_{11}}$	0.9218	1.3142	354.9318	354.9318	0.1647	0.1928	338.4184
$\hat{\bar{Y}}_{p_{12}}$	0.9712	1.5491	367.7088	367.7088	0.0787	0.1154	357.8285
$\hat{\bar{Y}}_{p_{13}}$	0.9975	1.6810	375.0921	375.0921	0.0290	0.0475	371.0234
$\hat{\bar{Y}}_{p_{14}}$	0.5000	0.0105	304.1857	304.1857	0.5060	0.0001	304.1748
$\hat{\bar{Y}}_{p_{15}}$	0.9974	1.6803	375.0531	375.0531	0.0293	0.0479	370.9498
$\hat{\bar{Y}}_{p_{16}}$	0.5132	0.0124	304.1895	304.1895	0.5060	0.0002	304.1748
$\hat{\bar{Y}}_{p_{17}}$	0.8636	1.0586	341.7007	341.7007	0.2537	0.2196	322.8924
$\hat{\bar{Y}}_{p_{18}}$	0.9843	1.6144	371.3460	371.3460	0.0542	0.0841	364.1476
$\hat{\bar{Y}}_{p_{20}}$	0.9940	1.6634	374.1000	374.1000	0.0357	0.0576	369.1661

Table 7 continued

Estimator	θ_i	$B\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\hat{\bar{Y}}_{p_{21}}$	0.9877	1.6314	372.2986	372.2986	0.0478	0.0752	365.8606
$\hat{\bar{Y}}_{p_{22}}$	0.8318	0.9292	335.3364	335.3364	0.2965	0.2132	317.0819
$\hat{\bar{Y}}_{p_{23}}$	0.9871	1.6281	372.1130	372.1130	0.0491	0.0769	365.5250
$\hat{\bar{Y}}_{p_{24}}$	0.8391	0.9582	336.7384	336.7384	0.2871	0.2157	318.2694
$\hat{\bar{Y}}_{p_{25}}$	0.8825	1.1392	345.7872	345.7872	0.2262	0.2171	327.1910
$\hat{\bar{Y}}_{p_{26}}$	0.9815	1.6000	370.5415	370.5415	0.0597	0.0913	362.7196
$\hat{\bar{Y}}_{p_{27}}$	0.9961	1.6737	374.6820	374.6820	0.0318	0.0517	370.2524
$\hat{\bar{Y}}_{p_{28}}$	0.6109	0.2194	307.3971	307.3971	0.4844	0.0360	304.3128
$\hat{\bar{Y}}_{p_{29}}$	0.9959	1.6726	374.6210	374.6210	0.0322	0.0523	370.1382
$\hat{\bar{Y}}_{p_{30}}$	0.6233	0.2505	308.2092	308.2092	0.4790	0.0446	304.3911
$\hat{\bar{Y}}_{p_{31}}$	0.9985	1.6862	375.3879	375.3879	0.0270	0.0444	371.5822
$\hat{\bar{Y}}_{p_{32}}$	0.3716	0.1715	309.4927	309.4927	0.4703	0.0577	304.5507
$\hat{\bar{Y}}_{p_{33}}$	0.9984	1.6858	375.3647	375.3647	0.0272	0.0447	371.5383
$\hat{\bar{Y}}_{p_{34}}$	0.3839	0.1609	308.5583	308.5583	0.4766	0.0482	304.4302
$\hat{\bar{Y}}_{p_{35}}$	0.9498	1.4452	361.9937	361.9937	0.1172	0.1563	348.6100
$\hat{\bar{Y}}_{p_{36}}$	0.9546	1.4682	363.2505	363.2505	0.1087	0.1482	350.5627

Table 8. Constant, Bias and Mean squared error of the Existing and Proposed estimators for Population 4

Estimator	θ_i	$B\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\bar{y}_{srs}$	-	-	448.5780	448.5780	-	-	-
$\hat{\bar{Y}}_R$	-	0.2542	18.3619	18.3619	-	-	-
$\hat{\bar{Y}}_{p_1}$	0.9903	0.2144	17.7773	17.7773	0.9311	0.0121	16.2363
$\hat{\bar{Y}}_{p_2}$	0.9321	0.0082	16.2333	16.2333	0.9344	0.0000	16.2307
$\hat{\bar{Y}}_{p_3}$	0.9786	0.1676	17.1984	17.1984	0.9323	0.0076	16.2329
$\hat{\bar{Y}}_{p_4}$	0.9906	0.2156	17.7934	17.7934	0.9310	0.0122	16.2364
$\hat{\bar{Y}}_{p_5}$	0.4970	0.8423	110.9857	110.9857	0.7296	0.5790	36.9976
$\hat{\bar{Y}}_{p_6}$	0.6171	0.7587	66.0867	66.0867	0.8266	0.3451	21.9799
$\hat{\bar{Y}}_{p_7}$	0.9987	0.2488	18.2780	18.2780	0.9300	0.0159	16.2404
$\hat{\bar{Y}}_{p_8}$	0.9329	0.0055	16.2319	16.2319	0.9344	0.0000	16.2307
$\hat{\bar{Y}}_{p_9}$	0.9957	0.2364	18.0897	18.0897	0.9304	0.0145	16.2387
$\hat{\bar{Y}}_{p_{10}}$	0.9789	0.1686	17.2095	17.2095	0.9323	0.0076	16.2330
$\hat{\bar{Y}}_{p_{11}}$	0.9901	0.2137	17.7674	17.7674	0.9311	0.0120	16.2362
$\hat{\bar{Y}}_{p_{12}}$	0.9907	0.2161	17.7996	17.7996	0.9310	0.0122	16.2364
$\hat{\bar{Y}}_{p_{13}}$	0.9999	0.2538	18.3558	18.3558	0.9298	0.0165	16.2412
$\hat{\bar{Y}}_{p_{14}}$	0.5000	0.8416	109.6730	109.6730	0.7324	0.5732	36.4262
$\hat{\bar{Y}}_{p_{15}}$	0.9998	0.2536	18.3520	18.3520	0.9298	0.0165	16.2411
$\hat{\bar{Y}}_{p_{16}}$	0.6200	0.7553	65.1890	65.1890	0.8286	0.3397	21.7747
$\hat{\bar{Y}}_{p_{17}}$	0.9687	0.1288	16.8141	16.8141	0.9331	0.0046	16.2315
$\hat{\bar{Y}}_{p_{18}}$	0.9971	0.2423	18.1775	18.1775	0.9302	0.0152	16.2395
$\hat{\bar{Y}}_{p_{19}}$	0.9309	0.0125	16.2366	16.2366	0.9344	0.0000	16.2307

Table 8 continued

Estimator	θ_i	$B\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_i\right)$	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_L & α_u	θ_{p_i} at α_a	$\text{Bias}\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a	$MSE\left(\hat{\bar{Y}}_{p_i}\right)$ at α_a
$\hat{\bar{Y}}_{p_{20}}$	0.9987	0.2490	18.2805	18.2805	0.9300	0.0160	16.2405
$\hat{\bar{Y}}_{p_{21}}$	0.9993	0.2513	18.3169	18.3169	0.9299	0.0162	16.2408
$\hat{\bar{Y}}_{p_{22}}$	0.8812	0.1815	17.6298	17.6298	0.9314	0.0109	16.2353
$\hat{\bar{Y}}_{p_{23}}$	0.9989	0.2495	18.2887	18.2887	0.9299	0.0160	16.2405
$\hat{\bar{Y}}_{p_{24}}$	0.9237	0.0383	16.2874	16.2874	0.9343	0.0004	16.2307
$\hat{\bar{Y}}_{p_{25}}$	0.9782	0.1661	17.1814	17.1814	0.9323	0.0074	16.2328
$\hat{\bar{Y}}_{p_{26}}$	0.9958	0.2369	18.0976	18.0976	0.9304	0.0145	16.2388
$\hat{\bar{Y}}_{p_{27}}$	0.9998	0.2533	18.3483	18.3483	0.9298	0.0165	16.2411
$\hat{\bar{Y}}_{p_{28}}$	0.6902	0.6531	45.7570	45.7570	0.8706	0.2153	18.2472
$\hat{\bar{Y}}_{p_{29}}$	0.9997	0.2528	18.3398	18.3398	0.9298	0.0164	16.2410
$\hat{\bar{Y}}_{p_{30}}$	0.7842	0.4563	27.3969	27.3969	0.9103	0.0851	16.5191
$\hat{\bar{Y}}_{p_{31}}$	0.9999	0.2538	18.3560	18.3560	0.9298	0.0165	16.2412
$\hat{\bar{Y}}_{p_{32}}$	0.4924	0.8433	112.9917	112.9917	0.7253	0.5877	37.8862
$\hat{\bar{Y}}_{p_{33}}$	0.9999	0.2536	18.3523	18.3523	0.9298	0.0165	16.2411
$\hat{\bar{Y}}_{p_{34}}$	0.6127	0.7637	67.4673	67.4673	0.8237	0.3534	22.3027
$\hat{\bar{Y}}_{p_{35}}$	0.9844	0.1909	17.4704	17.4704	0.9317	0.0097	16.2343
$\hat{\bar{Y}}_{p_{36}}$	0.9941	0.2299	17.9954	17.9954	0.9306	0.0138	16.2379

From the values of Table 5—Table 8, it is observed that the bias of the proposed modified ratio estimator $\hat{\bar{Y}}_{p_j}; j = 1,2,\ldots,36$ is less than the bias of the usual ratio estimator and the existing modified ratio estimators $\hat{\bar{Y}}_i; i = 1,2,3,\ldots,36$. Similarly, the mean squared error of the proposed modified ratio estimator $\hat{\bar{Y}}_{p_j}; j = 1,2,\ldots,36$

is less than the variance of SRSWOR sample mean, the mean squared error of the usual ratio estimator and the existing modified ratio estimators $\hat{\bar{Y}}_{p_j}; j = 1, 2, \ldots, 36$ for all four populations.

Conclusion

In this article, a generalized modified ratio estimator has been suggested using the known population parameters of the auxiliary variable. Moreover, many modified ratio estimators have been introduced in this article, and have not been discussed earlier in the literature. The bias and mean squared error of the proposed generalized modified ratio estimator are obtained. Furthermore, the conditions have been derived for which the proposed estimator is more efficient than the existing estimators, and it is shown that the SRSWOR sample mean, the usual ratio estimator, the linear regression and the existing modified ratio estimators are particular cases of the proposed estimator. The performances of the proposed estimator are also assessed for some known populations. It is observed that the bias and the mean squared errors of the proposed estimators are less than the bias and the mean squared error of the existing estimators. Moreover, the proposed estimator will be a generalized modified ratio estimator for estimating the population mean of the study variable using the known population parameters of the auxiliary variable.

Acknowledgements

The author wishes to record his gratitude and thanks to University Grants Commission (UGC) for the financial assistance through UGC-Major Research Project.

References

Abdia, Y. Z., & Shahbaz, M. Q. (2006). A comparative study of generalized ratio and regression estimators with their classical counterparts. *Pakistan Journal of Statistics and Operation Research, 2*(1), 63-68.

Ahmad, Z., Hanif, M., & Ahmad, M. (2009). Generalized regression-cum-ratio estimators for two-phase sampling using multi-auxiliary variables. *Pakistan Journal of Statistics*, *25*(2), 93-106.

Al-jararha, J., & Al-Haj Ebrahem, M. (2012). A ratio estimator under general sampling design. *Austrian Journal of Statistics, 41*(2), 105-115.

Bhushan, S. (2012). Some efficient sampling strategies based on ratio-type estimator. *Electronic Journal of Applied Statistical Analysis, 5*(1), 74-88.

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition, Wiley Eastern Limited.

Dalabehera, M. & Sahoo, L. N. (1994). Comparison of six almost unbiased ratio estimators. *QUESTIIO, 18*(3), 369-375.

David, I. P., & Sukhatme, B .V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association, Theory and Methods Section, 69*(346): 464-466.

Goodman, L. A., & Hartley, H. O. (1958). The precision of unbiased ratio-type estimators. *Journal of the American Statistical Association, 53*(282): 491-508.

Gupta, S., & Shabbir, J. (2008). On improvement in estimating the population mean in simple random sampling. *Journal of Applied Statistics, 35*(5), 559–566.

Jhajj, H.S., Sharma, M. K., & Grover, L.K. (2006). Dual of ratio estimators of finite population mean obtained on using linear transformation to auxiliary variable. *Journal of Japan Statistical Society, 36*(1), 107–119.

Kadilar, C., & Cingi, H. (2003). A study on the chain ratio type estimator. *Hacettepe Journal of Mathematics and Statistics,* Vol. 32, 105-108.

Kadilar, C., & Cingi, H. (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation, 151*, 893-902.

Kadilar, C., & Cingi, H. (2006a). An improvement in estimating the population mean by using the correlation co-efficient. *Hacettepe Journal of Mathematics and Statistics, 35*(1), 103-109.

Kadilar, C., & Cingi, H. (2006b). Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters, 19*, 75-79.

Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N., & Smarandache, F. (2007). A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East Journal of Theoretical Statistics, 22*, 181-191.

Koyuncu, N., & Kadilar, C. (2009). Efficient Estimators for the Population Mean. *Hacettepe Journal of Mathematics and Statistics, 38*(2), 217-225.

Kulkarni, S. P. (1978). A note on modified ratio estimator using transformation. *Journal of the Indian Society of Agricultural Statistics*, *30*(2), 125–128.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

Naik, V. D., & Gupta, P. C. (1991): A general class of estimators for estimating population mean using auxiliary information. *Metrika*, *38*, 11–17.

Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, *45*, 154-165.

Pathak, P.K. (1964). On sampling schemes providing unbiased ratio estimators. *The Annals of Mathematical Statistics*, *35*(1), 222-231.

Perri, P. F. (2007). Improved ratio-cum-product type estimators. *Statistics in Transition*, *8*(1), 51-69.

Ray, S. K., & Sahai, A. (1980). Efficient families of ratio and product-type estimators. *Biometrika, 67*, 211–215.

Reddy, V. N. (1973). On ratio and product methods of estimation. *Sankhya B*, *35*(3), 307-316.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association, 82* (399), 826-831.

Sen, A. R. (1993). Some early developments in ratio estimation. *Biometrical Journal*, *35*(1), 3-13.

Shabbir, J., & Yaab, M. Z. (2003). Improvement over transformed auxiliary variable in estimating the finite populations mean. *Biometrical Journal, 45*(6), 723–729.

Sharma, B., & Tailor, R. (2010). A new ratio-cum-dual to ratio estimator of finite population mean in simple random sampling. *Global Journal of Science Frontier Research, 10*(1), 27-31.

Singh, D., & Chaudhary, F. S. (1986). *Theory and analysis of sample survey designs*. New Delhi: New Age International Publisher.

Singh, G. N. (2003). On the improvement of product method of estimation in sample surveys. *Journal of the Indian Society of Agricultural Statistics, 56*(3), 267–275.

Singh, H. P. & Espejo, M. R. (2003). On linear regression and ratio-product estimation of a finite population mean. *The Statistician*, *52*, 59-67.

Singh, H. P., & Agnihotri, N. (2008). A general procedure of estimating population mean using auxiliary information in sample surveys. *Statistics in Transition*, *9*(1), 71–87.

Singh, H. P., & Solanki, R. S. (2012). An alternative procedure for estimating the population mean in simple random sampling. *Pakistan Journal of Statistics and Operation Research*, *8*(2), 213-232.

Singh, H. P., & Tailor, R. (2003). Use of known correlation co-efficient in estimating the finite population means. *Statistics in Transition, 6* (4), 555-560.

Singh, H. P., & Tailor, R. (2005). Estimation of finite population mean with known co-efficient of variation of an auxiliary. *Statistica*, anno LXV, n.3, pp 301-313.

Singh, H. P., Tailor, R., Singh. S., & Kim, J. M. (2008). A modified estimator of population mean using power transformation. *Statistical Papers*, *49*, 37–58.

Singh, H. P., Tailor, R., Tailor, R. and Kakran, M. S. (2004): An Improved Estimator of Population Mean Using Power Transformation. *Journal of the Indian Society of Agricultural Statistics*, *58*(2), 223-230.

Sisodia, B. V. S., & Dwivedi, V. K. (1981). A modified ratio estimator using co-efficient of variation of auxiliary variable. *Journal of the Indian Society of Agricultural Statistics*, *33*(2), 13-18.

Solanki, R. S., Singh, H. P., & Rathour, A. (2012). An alternative estimator for estimating the finite population mean using auxiliary information in sample surveys. *ISRN Probability and Statistics*, Article ID 657682, 14 pp.

Srivenkataramana, T. (1980). A dual to ratio estimator in sample surveys. Biometrika, *37*, 199–204.

Subramani, J., & Kumarapandiyan, G. (2012a). Modified ratio estimators using known median and co-efficient of kurtosis. *American Journal of Mathematics and Statistics*, *2*(4), 95-100.

Subramani, J., & Kumarapandiyan, G. (2012b). Estimation of population mean using known median and co-efficient of skewness. *American Journal of Mathematics and Statistics*. *2*(5), 101-107.

Subramani, J., & Kumarapandiyan, G. (2012c). Estimation of population mean using co-efficient of variation and median of an auxiliary variable, *International Journal of Probability and Statistics*, *1*(4), 111-118.

Subramani, J., & Kumarapandiyan, G. (2013a). A new modified ratio estimator for estimation of population mean when median of the auxiliary

variable is known. *Pakistan Journal of Statistics and Operation Research, 9*(2), 137-145.

Subramani, J., & Kumarapandiyan, G. (2013b). Estimation of Population Mean Using Known Correlation Co-Efficient and Median. *Journal of Statistical Theory and Applications* (to appear).

Tailor, R., & Sharma, B. (2009). A modified ratio-cum-product estimator of finite population mean using known coefficient of variation and coefficient of kurtosis. *Statistics in Transition-New Series, 10*(1), 15-24.

Tin, M. (1965). Comparison of some ratio estimators. *Journal of the American Statistical Association, 60*, 294-307.

Upadhyaya, L. N., & Singh, H.P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, V*41*(5), 627-636.

Yan, Z., & Tian, B. (2010). Ratio method to the mean estimation using co-efficient of skewness of auxiliary variable. *ICICA 2010*, Part II, CCIS 106, pp. 103–111.

Appendix A

An expression for the bias and mean squared error of the proposed estimators $\hat{\bar{Y}}_{p_j}; i = 1,2,3,\ldots,36$ was derived to first order of approximation with the following notations:

Let us define $e_0 = \dfrac{\bar{y}-\bar{Y}}{\bar{Y}}$ and $e_1 = \dfrac{\bar{x}-\bar{X}}{\bar{X}}$. Further, $\bar{y} = \bar{Y}(1+e_0)$ and $\bar{x} = \bar{X}(1+e_1)$ and from the definition of $e_{0 \text{ and }} e_1$:

$$E[e_0] = E[e_1] = 0$$

$$E\left[e_0^2\right] = \frac{(1-f)}{n} C_y^2$$

$$E\left[e_1^2\right] = \frac{(1-f)}{n} C_x^2$$

$$E[e_0 e_1] = \frac{(1-f)}{n} \rho C_y C_x \text{ where } C_x = \frac{S_x}{\bar{X}}, C_y = \frac{S_y}{\bar{Y}} \text{ and } \rho = \frac{S_{xy}}{S_x S_y}$$

The bias of a class of proposed estimators $\hat{\bar{Y}}_{p_i}; i = 1,2,3,\ldots,36$ is derived and is:

$$\hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(\bar{x} + (1+\alpha)\lambda_i\right)}\left(\bar{X} + (1+\alpha)\lambda_i\right); i = 1,2,3,\ldots,36$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(\bar{X} + e_1\bar{X} + (1+\alpha)\lambda_i\right)}\left(\bar{X} + (1+\alpha)\lambda_i\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(\bar{X} + (1+\alpha)\lambda_i\right)\left(1 + \dfrac{e_1\bar{X}}{\left(\bar{X} + (1+\alpha)\lambda_i\right)}\right)}\left(\bar{X} + (1+\alpha)\lambda_i\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(1+\theta_{p_i} e_1\right)} \text{ where } \theta_{p_i} = \frac{\bar{X}}{\bar{X} + (1+\alpha)\lambda_i}$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{y}\left(1+\theta_{p_i} e_1\right)^{-1}$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{y}\left(1 - \theta_{p_i} e_1 + \theta_{p_i}^2 e_1^2 - \theta_{p_i}^3 e_1^3 + \ldots\right)$$

Neglecting the terms more than 2nd order, results in

$$\hat{\bar{Y}}_{p_i} = \bar{y}\left(1-\theta_{p_i}e_1+\theta_{p_i}^2e_1^2\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \left(\bar{Y}\left(1+e_0\right)\right)\left(1-\theta_{p_i}e_1+\theta_{p_i}^2e_1^2\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \left(\bar{Y}+\bar{Y}e_0\right)\left(1-\theta_{p_i}e_1+\theta_{p_i}^2e_1^2\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{Y}+\bar{Y}e_0-\bar{Y}\theta_{p_i}e_1-\bar{Y}\theta_{p_i}e_0e_1+\bar{Y}\theta_{p_i}^2e_1^2+\bar{Y}\theta_{p_i}^2e_0e_1^2$$

Neglecting the terms more than 3rd order, results in

$$\hat{\bar{Y}}_{p_i} = \bar{Y}+\bar{Y}e_0-\bar{Y}\theta_{p_i}e_1-\bar{Y}\theta_{p_i}e_0e_1+\bar{Y}\theta_{p_i}^2e_1^2$$

$$\Rightarrow \hat{\bar{Y}}_{p_i}-\bar{Y} = \bar{Y}e_0-\bar{Y}\theta_{p_i}e_1-\bar{Y}\theta_{p_i}e_0e_1+\bar{Y}\theta_{p_i}^2e_1^2$$

Taking expectation on both sides, results in

$$E\left(\hat{\bar{Y}}_{p_i}-\bar{Y}\right) = \bar{Y}E\left(e_0\right)-\bar{Y}\theta_{p_i}E\left(e_1\right)-\bar{Y}\theta_{p_i}E\left(e_0e_1\right)+\bar{Y}\theta_{p_i}^2E\left(e_1^2\right)$$

$$\Rightarrow \text{Bias}\left(\hat{\bar{Y}}_{p_i}\right) = \bar{Y}\theta_{p_i}^2E\left(e_1^2\right)-\bar{Y}\theta_{p_i}E\left(e_0e_1\right)$$

$$\Rightarrow \text{Bias}\left(\hat{\bar{Y}}_{p_i}\right) = \bar{Y}\theta_{p_i}^2\frac{\left(1-f\right)}{n}C_x^2-\bar{Y}\theta_{p_i}\frac{\left(1-f\right)}{n}\rho C_yC_x$$

$$\Rightarrow \text{Bias}\left(\hat{\bar{Y}}_{p_i}\right) = \frac{\left(1-f\right)}{n}\left(\bar{Y}\theta_{p_i}^2C_x^2-\bar{Y}\theta_{p_i}\rho C_yC_x\right)$$

$$\Rightarrow \text{Bias}\left(\hat{\bar{Y}}_{p_i}\right) = \frac{\left(1-f\right)}{n}\bar{Y}\left(\theta_{p_i}^2C_x^2-\theta_{p_i}\rho C_yC_x\right) \text{ where } \theta_{p_i} == \frac{\bar{X}}{\bar{X}+\left(1+\alpha\right)\lambda_i}$$

The mean squared error of the proposed estimator $\hat{\bar{Y}}_{p_i}; i = 1,2,3,\ldots,36$ to first order of approximation is derived and is:

$$\hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(\bar{x}+\left(1+\alpha\right)\lambda_i\right)}\left(\bar{X}+\left(1+\alpha\right)\lambda_i\right); i = 1,2,3,\ldots,36$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(\bar{X}+e_1\bar{X}+\left(1+\alpha\right)\lambda_i\right)}\left(\bar{X}+\left(1+\alpha\right)\lambda_i\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(\bar{X}+\left(1+\alpha\right)\lambda_i\right)\left(1+\frac{e_1\bar{X}}{\left(\bar{X}+\left(1+\alpha\right)\lambda_i\right)}\right)}\left(\bar{X}+\left(1+\alpha\right)\lambda_i\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \frac{\bar{y}}{\left(1+\theta_{p_i}e_1\right)} \text{ where } \theta_{p_i} = \frac{\bar{X}}{\bar{X}+\left(1+\alpha\right)\lambda_i}$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{y}\left(1+\theta_{p_i}e_1\right)^{-1}$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{y}\left(1-\theta_{p_i}e_1+\theta_{p_i}^2e_1^2-\theta_{p_i}^3e_1^3+\ldots\right)$$

Neglecting the terms more than 1st order, results in

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{y}\left(1-\theta_{p_i}e_1\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \left(\bar{Y}\left(1+e_0\right)\right)\left(1-\theta_{p_i}e_1\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \left(\bar{Y}+\bar{Y}e_0\right)\left(1-\theta_{p_i}e_1\right)$$

$$\Rightarrow \hat{\bar{Y}}_{p_i} = \bar{Y}+\bar{Y}e_0-\bar{Y}\theta_{p_i}e_1-\bar{Y}\theta_{p_i}e_0e_1$$

$$\Rightarrow \hat{\bar{Y}}_{p_i}-\bar{Y} = \bar{Y}e_0-\bar{Y}\theta_{p_i}e_1-\bar{Y}\theta_{p_i}e_0e_1$$

Squaring both sides

$$\Rightarrow \left(\hat{\bar{Y}}_{p_i}-\bar{Y}\right)^2 = \left(\bar{Y}e_0-\bar{Y}\theta_{p_i}e_1-\bar{Y}\theta_{p_i}e_0e_1\right)^2$$

Neglecting the terms more than 2nd order, results in

$$\left(\hat{\bar{Y}}_{p_i}-\bar{Y}\right)^2 = \bar{Y}^2e_0{}^2-\bar{Y}^2\theta_{p_i}^2e_1^2-2\bar{Y}^2\theta_{p_i}e_0e_1$$

Taking expectation on both sides results in:

$$E\left(\hat{\bar{Y}}_{p_i}-\bar{Y}\right)^2 = E\left(\bar{Y}^2e_0^2\right)+\bar{Y}^2\theta_{p_i}^2E\left(e_1^2\right)-2\bar{Y}^2\theta_{p_i}E\left(e_0e_1\right)$$

$$\Rightarrow MSE\left(\hat{\bar{Y}}_{p_i}\right) = \frac{\left(1-f\right)}{n}\left(\bar{Y}^2C_y^2+\bar{Y}^2\theta_{p_i}^2C_x^2-2\bar{Y}^2\theta_{p_i}\rho C_yC_x\right); i=1,2,3,\ldots,36$$

$$\text{where } \theta_{p_i} = \frac{\bar{X}}{\bar{X}+\left(1+\alpha\right)\lambda_i}$$

Appendix B

The conditions for which proposed estimator $\hat{\bar{Y}}_{p_i}$ perform better than the SRSWOR sample mean are derived and are given below:

For $MSE\left(\hat{\bar{Y}}_{p_j}\right) \le V\left(\bar{y}_r\right)$

$$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\right)\le\frac{(1-f)}{n}S_y^2$$
$$\Rightarrow\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\right)\le\frac{(1-f)}{n}\bar{Y}^2C_y^2$$
$$\Rightarrow\left(C_y^2+\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\right)\le C_y^2$$
$$\Rightarrow\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\le C_y^2$$
$$\Rightarrow\theta_{p_i}^2C_x^2\le 2\rho\theta_{p_i}C_xC_y$$
$$\Rightarrow\theta_{p_i}C_x\le 2\rho C_y$$
$$\Rightarrow\theta_{p_i}\le 2\rho\frac{C_y}{C_x}$$

That is, $MSE\left(\hat{\bar{Y}}_{p_i}\right)\le V\left(\bar{y}_r\right)$ if $\theta_{p_i}\le 2\rho\frac{C_y}{C_x}$

Appendix C

The conditions for which proposed estimator $\hat{\bar{Y}}_{p_i}$ perform better than the usual ratio estimator are derived and are given below:

For $MSE\left(\hat{\bar{Y}}_{p_j}\right)\le MSE\left(\hat{\bar{Y}}_R\right)$

$$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\right)\le\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2+C_x^2-2\rho C_xC_y\right)$$
$$\Rightarrow\left(C_y^2+\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\right)\le\left(C_y^2+C_x^2-2\rho C_xC_y\right)$$
$$\Rightarrow\theta_{p_i}^2C_x^2-2\rho\theta_{p_i}C_xC_y\le C_x^2-2\rho C_xC_y$$
$$\Rightarrow\theta_{p_i}^2C_x^2-C_x^2-2\rho\theta_{p_i}C_xC_y+2\rho C_xC_y\le 0$$
$$\Rightarrow\left(\theta_{p_i}^2-1\right)C_x^2-2\rho C_xC_y\left(\theta_p-1\right)\le 0$$
$$\Rightarrow\left(\theta_{p_i}-1\right)\left(\left(\theta_{p_i}+1\right)C_x^2-2\rho C_xC_y\right)\le 0$$

Condition 1: $\left(\theta_{p_i}-1\right)\le 0$ and $\left(\left(\theta_{p_i}+1\right)C_x^2-2\rho C_xC_y\right)\ge 0$

$$\Rightarrow \theta_{p_i} \le 1 \text{ and } \left(\theta_{p_i}+1\right)C_x^2 \ge 2\rho C_x C_y$$

$$\Rightarrow \theta_{p_i} \le 1 \text{ and } \theta_{p_i} \ge \frac{2\rho C_y}{C_x} - 1$$

$$\Rightarrow \frac{2\rho C_y}{C_x} - 1 \le \theta_{p_i} \le 1$$

Condition 2: $\left(\theta_{p_i}-1\right) \ge 0$ and $\left(\left(\theta_{p_i}+1\right)C_x^2 - 2\rho C_x C_y\right) \le 0$

$$\Rightarrow \theta_{p_i} \ge 1 \text{ and } \left(\theta_{p_i}+1\right)C_x^2 - 2\rho C_x C_y$$

$$\Rightarrow \theta_{p_i} \ge 1 \text{ and } \theta_{p_i} \le \frac{2\rho C_y}{C_x} - 1$$

$$\Rightarrow 1 \le \theta_{p_i} \le \frac{2\rho C_y}{C_x} - 1$$

That is, $MSE\left(\hat{\bar{Y}}_{p_i}\right) \le MSE\left(\hat{\bar{Y}}_R\right)$ either $\frac{2\rho C_y}{C_x} - 1 \le \theta_{p_i} \le 1$ (or) $1 \le \theta_{p_i} \le \frac{2\rho C_y}{C_x} - 1$

Appendix D

The conditions for which proposed estimator $\hat{\bar{Y}}_{p_i}$ perform better than the existing modified ratio estimators (Class 1) are derived and are given below:

For $MSE\left(\hat{\bar{Y}}_{p_j}\right) \le MSE\left(\hat{\bar{Y}}_i\right); i = 1, 2, 3, \ldots, 36$

$$\frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_{p_i}^2 C_x^2 - 2\rho\theta_{p_i} C_x C_y\right) \le \frac{(1-f)}{n}\bar{Y}^2\left(C_y^2 + \theta_i^2 C_x^2 - 2\rho\theta_i C_x C_y\right)$$

$$\Rightarrow \left(C_y^2 + \theta_{p_i}^2 C_x^2 - 2\rho\theta_{p_i} C_x C_y\right) \le \left(C_y^2 + \theta_i^2 C_x^2 - 2\rho\theta_i C_x C_y\right)$$

$$\Rightarrow \theta_{p_i}^2 C_x^2 - 2\rho\theta_{p_i} C_x C_y \le \theta_i^2 C_x^2 - 2\rho\theta_i C_x C_y$$

$$\Rightarrow \theta_{p_i}^2 C_x^2 - \theta_i^2 C_x^2 - 2\rho\theta_{p_i} C_x C_y + 2\rho\theta_i C_x C_y \le 0$$

$$\Rightarrow \left(\theta_{p_i}^2 - \theta_i^2\right)C_x^2 - 2\rho\theta_{p_i} C_x C_y\left(\theta_{p_i} - \theta_i\right) \le 0$$

$$\Rightarrow \left(\theta_{p_i} - \theta_i\right)\left(\left(\theta_{p_i} + \theta_i\right)C_x^2 - 2\rho C_x C_y\right) \le 0$$

Condition 1: $\left(\theta_{p_i} - \theta_i\right) \le 0$ and $\left(\left(\theta_p + \theta_i\right)C_x^2 - 2\rho C_x C_y\right) \ge 0$

$$\Rightarrow \theta_{p_i} \le \theta_i \text{ and } \left(\theta_{p_i} + \theta_i\right) C_x^2 \ge 2\rho C_x C_y$$

$$\Rightarrow \theta_{p_i} \le \theta_i \text{ and } \theta_{p_i} \ge \frac{2\rho C_y}{C_x} - \theta_i$$

$$\Rightarrow \frac{2\rho C_y}{C_x} - \theta_i \le \theta_{p_i} \le \theta_i$$

Condition 2: $\left(\theta_{p_i} - \theta_i\right) \ge 0$ and $\left(\left(\theta_{p_i} + \theta_i\right) C_x^2 - 2\rho C_x C_y\right) \le 0$

$$\Rightarrow \theta_{p_i} \ge \theta_i \text{ and } \left(\theta_{p_i} + \theta_i\right) C_x^2 - 2\rho C_x C_y$$

$$\Rightarrow \theta_{p_i} \ge \theta_i \text{ and } \theta_{p_i} \le \frac{2\rho C_y}{C_x} - \theta_i$$

$$\Rightarrow \theta_i \le \theta_{p_i} \le \frac{2\rho C_y}{C_x} - \theta_i$$

That is, $MSE\left(\hat{\bar{Y}}_{p_i}\right) \le MSE\left(\hat{\bar{Y}}_i\right)$ either

$\frac{2\rho C_y}{C_x} - \theta_i \le \theta_{p_i} \le \theta_i$ (or) $\theta_i \le \theta_{p_i} \le \frac{2\rho C_y}{C_x} - \theta_i$

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 156-170.

ISSN 1538 − 9472

Intrinsically Ties Adjusted Non-Parametric Method for the Analysis of Two Sampled Data

G. U. Ebuh
Nnamdi Azikiwe University
Awka, Nigeria

I. C. A. Oyeka
Nnamdi Azikiwe University
Awka, Nigeria

A non-parametric method for the analysis of two sample data is proposed that intrinsically and structurally adjusts the test statistic for the possible presence of tied observations between the sampled populations, thereby obviating the need to require the populations to be continuous. The populations may be measurements on as low as the ordinal scale, and need not be homogeneous. In cases where the null hypotheses are rejected, the test statistic enables the determination of which of the sampled populations is likely to be responsible for the rejection (a determination which the Wilcoxon Mann Whitney test cannot handle). The proposed method is illustrated with some data, and shown to compare favorably with some existing methods available for the same purpose.

Keywords: Two sampled data, proposed method, observations, hypothesis.

Introduction

Suppose a researcher has collected two random samples of sizes n_1 and n_2 from two populations x_1 and x_2 respectively. This researcher may be interested in testing a null hypothesis such as H_0: $\mu_2 = \frac{\alpha}{\beta}\mu_1 + \sigma$ verses either a two sided or any of the one sided alternative hypotheses, where α and β are non-zero real numbers, and σ is any real number including zero. The null hypothesis is that one of the populations is on the average at least (at most) a multiple (a proportion) of the other population. This situation may arise in many cases. In the health delivery system, researchers may be interested in testing the hypothesis that the effective dosage of a certain treatment drug is at least c times that of a control

Dr. Ebuh is a Senior Lecturer in the Department of Statistics, Faculty of Physical Sciences. Email him at: ablegod007@yahoo.com. Professor Oyeka is Professor of Statistics in the Department of Statistics, Faculty of Physical Sciences.

drug, where $c = \frac{\alpha}{\beta}$, or that the bed occupancy rate of public hospitals is at most *c* times that of private hospitals. In business studies, interest may be in determining whether the cost of a certain line of products in a certain retail shop or market is at least *c* times higher on the average than the cost in another retail shop or market, or whether Gender B workers on the average earn at most *c* times that of their Gender A counterparts of equal skill. In education and public affairs, interest may be in whether students of a certain instructor, or candidates under a certain panel of judges, score at least *c* times more than students of another instructor or candidates under another panel of judges; or whether the rate at which a certain set of trial judges deliver judgment in cases is at most *c* times the rate at which a second set of trial judges deliver judgments during the year, etc.

In each of these and similar situations, the parametric *t* test cannot properly be used to test the hypothesis without first using appropriate data transformation, given the problem of homogeneity. If *c* = 1, then the *t* test may be used to test the desired null hypothesis provided the sampled populations are mutually homogeneous and normally distributed. If *c* is any real number other than 1, then the *t* test cannot be properly used for data analysis, even if the populations are normally distributed, without first applying some appropriate transformation to ensure homogeneity. This is because multiplying or dividing a data set by some non-zero constant changes the variance of the data set by the square of the constant, thereby violating the assumption of homogeneity necessary for the valid application of the *t* test.

Rather than applying some data transformation aimed at achieving homogeneity of variances, which may not be readily available, use of non-parametric statistical methods in these situations is usually preferable. If the sampled populations are related, paired or matched, then non-parametric methods that readily suggest themselves are the sign test and the Wilcoxon's Signed Rank Sum Test (Gibbons, 1971).

The problem with these two tests is that they require the sampled populations to be continuous, thereby theoretically making no definite provisions for the possible presence of tied observations between the populations. Oyeka et al. (2009) developed a method for the analysis of these types of data that intrinsically and structurally adjusts the test statistic for the presence of any ties between the sampled populations, which may now be measurements on as low as the ordinal scale and need not be continuous. If the sampled populations are independent, then the non-parametric methods often used in their analysis include

the median test and the Wilcoxon Mann Whitney U. Test (Gibbons, 1971; Oyeka, 2009).

A problem with these two test statistics is that they often resolve the problem of ties between sampled populations by assigning tied observations their mean ranks. If the numbers of ties are large, this approach would tend to compromise the power of the tests, which may seriously affect any conclusions based on them. An alternative non-parametric statistical method is proposed to test the desired null hypothesis when the populations are independent. The proposed method intrinsically and structurally adjusts the test statistic for the possible presence of ties between the sampled populations, obviating the need to require the populations to be continuous. These populations may therefore be measurements on as low as the ordinal scale. Other authors who have done some research in this area include Afuecheta et al (2012), Ebuh & Oyeka (2012), & Ebuh et al (2012).

The Proposed Method

Let x_{ij} be the ith observation independently drawn from population x_j, for $i = 1,2,\ldots,n_j$; $j = 1,2$. Population x_j may be measurements on as low as the ordinal scale and need not be continuous. To develop the test statistic, first list unchanged all the observations x_i from one of the sampled populations x_1 while multiplying, (or dividing) each of the observations x_{j2} from the other sampled population x_2 by the constant $c = \frac{\alpha}{\beta}$, then add (or subtract) the constant $d = \sigma$ before pooling them together. For the purpose of determining the common median the pooled observations are then ranked together, either from the smallest to the largest or the largest to the smallest. Tied observations are assigned their mean ranks. The common median *Mc* of the pooled sample observations is then determined. Let r_{ij} be the rank assigned to x_{ij} in the combined ranking of $n=n_1+n_2$ sample observations. The proposed method is developed based on the common median *Mc* in particular. For the purposes of comparison with the Wilcoxon Mann Whitney test, let r_{i1} be the rank assigned x_{i1} and r_{j2} the rank assigned to x_{j2}, adjusted values of x_{j2} in the combined ranking of the 'n' = n_1 + n_2 sample observations; for $i = 1,2,\ldots,n_1$ and $j = 1,2,\ldots n_2$.

Let

$$u_{i1} = \begin{cases} 1, \ if \ x_{i1} > Mc \\ 0, \ if \ x_{i1} = Mc \\ -1, if \ x_{i1} < Mc \end{cases} \tag{1}$$

for $i = 1,2\ldots n_{1.}$ Let

$$\overset{+}{\pi}_1 = P\left(u_{1i} = 1\right); \overset{0}{\pi}_1 = P\left(u_{1i} = 0\right); \overset{-}{\pi}_1 = P\left(u_{1i} = -1\right) \tag{2}$$

where $\overset{+}{\pi}_1 + \overset{0}{\pi}_1 + \overset{-}{\pi}_1 = 1$ (3)

and

$$W_1 = \sum_{i=1}^{n_1} \mu_{i1} \ . \tag{4}$$

W_1 is the difference between the numbers of sample observations in population X_1 that are greater than, and the number of sample observations in X_1 that are less than, the common median *Mc*.

Similarly, let

$$u_{i2} = \begin{cases} 1, \ if \ x_{j2} > Mc \\ 0, \ if \ x_{j2} = Mc \\ -1, if \ x_{j2} < Mc \end{cases} \tag{5}$$

for $j = 1,2,\ldots,n_2$. Let

$$\overset{+}{\pi}_2 = P\left(u_{j2} = 1\right); \overset{0}{\pi}_2 = P\left(u_{j2} = 0\right); \overset{-}{\pi}_2 = P\left(u_{j2} = -1\right) \ , \tag{6}$$

where $\overset{+}{\pi}_2 + \overset{0}{\pi}_2 + \overset{-}{\pi}_2 = 1$, and define (7)

$$W_2 = \sum_{j=1}^{n_2} \mu_{j2} \ . \tag{8}$$

W_2 is the difference between the numbers of sample observations in population X_2 that are greater than, and the number of sample observations in X_2 that are less than, the common median *Mc*. Thus,

$$\mathrm{E}\left(\mu_{i1}\right)=\overset{+}{\pi}_1-\overset{-}{\pi}_1;\ \mathrm{Var}(\mu_{i1})=\overset{+}{\pi}_1+\overset{-}{\pi}_1-\left(\overset{+}{\pi}_1-\overset{-}{\pi}_1\right)^2\ , \tag{9}$$

and

$$\mathrm{E}\left(\mu_{i1}\right)=n_1\left(\overset{+}{\pi}_1-\overset{-}{\pi}_1\right), \text{ and} \tag{10}$$

$\mathrm{Var}(W_1) = \sum_{i=1}^{n_1} \mathrm{Var}(\mu_{i1})$. Thus,

$$\mathrm{Var}(W_1)=n_1\left(\overset{+}{\pi}_1+\overset{-}{\pi}_1-\left(\overset{+}{\pi}_1-\overset{-}{\pi}_1\right)^2\right)\ . \tag{11}$$

Note $\overset{+}{\pi}_1, \overset{0}{\pi}_1 \text{ and } \overset{-}{\pi}_1$ are respectively the probabilities that a randomly selected observation from population x_1 is on the average greater (higher), the same as (equal to), or smaller (lower) than the common median *Mc* of the combined sample observations. Their sample estimates are respectively

$$\hat{\overset{+}{\pi}}_1=\frac{f_1^+}{n_1};\hat{\overset{0}{\pi}}_1=\frac{f_1^0}{n_1};\hat{\overset{-}{\pi}}_1=\frac{f_1^-}{n_1}\ , \tag{12}$$

where $f_1^+, f_1^0, \text{and} f_1^-$ are respectively the number of 1's, 0's and -1's in the frequency distribution of the n_1 values of these numbers in u_{ij}, i =1,2,…,n_1.

Note from Equation 4

$$\mathrm{W}_1 = f_1 - f_1^- \tag{13}$$

$$\mathrm{E}\left(\mu_{i2}\right)=\overset{+}{\pi}_2-\overset{-}{\pi}_2;\ \mathrm{Var}(\mu_{i2})=\overset{+}{\pi}_2+\overset{-}{\pi}_2-\left(\overset{+}{\pi}_2-\overset{-}{\pi}_2\right)^2 \tag{14}$$

$$\mathrm{E}\left(W_2\right)=n_2\left(\overset{+}{\pi}_2-\overset{-}{\pi}_2\right) \tag{15}$$

$$\mathrm{Var}(W_2)=n_2\left(\overset{+}{\pi}_2+\overset{-}{\pi}_2-\left(\overset{+}{\pi}_2-\overset{-}{\pi}_2\right)^2\right)\ . \tag{16}$$

Note $\overset{+}{\pi}_2, \overset{0}{\pi}_2 \text{ and } \overset{-}{\pi}_2$ are respectively the probabilities that a randomly selected adjusted sample observation from population x_2 is on the average greater

(higher), the same as (equal to), or smaller (lower) than the common median *Mc*. Their sample estimates are respectively

$$\hat{\pi}_2^{+} = \frac{f_2^{+}}{n_2}; \hat{\pi}_2^{0} = \frac{f_2^{0}}{n_2}; \hat{\pi}_2^{-} = \frac{f_2^{-}}{n_2}, \tag{17}$$

where $f_2^{+}, f_2^{0}, and f_2^{-}$ are respectively the number of 1's, 0's and -1's in the frequency distribution of these numbers in u_{j2}, i =1,2,…,n_2.
Note from Equation 8 that

$$W_2 = f_2^{+} + f_2^{-} \tag{18}$$

A null hypothesis that is usually of interest in two sample problems, particularly when non-parametric methods are used, is that the two populations have equal medians M_0. If population x_1 has median M_1 and population x_2 has median M_2, then a null hypothesis of interest may be

$$H_0 : M_1 = M_2 = M_0 \tag{19}$$

versus any desired alternative hypothesis.

Using the median test or the Wilcoxon Mann Whitney test to list the null hypothesis of Equation 17, and given the rejection of the null hypothesis, one could not immediately say which of the sampled populations actually led to the rejection of H_0. This is because one of the population medians may (or may not) be equal to the hypothesized value M_0 whereas the other population median may (or may not) be equal to M_0, but the test being used may not immediately reveal this pattern. To help determine this possibility, the null hypothesis of Equation 19 can be alternatively expressed as

$$H_{01} : M_1 = M_0 \text{versus } H_{11} : M_1 \neq M_0 \tag{20}$$

and

$$H_{02} : M_2 = M_0 \text{versus } H_{12} : M_2 \neq M_0 \tag{21}$$

If the null hypothesis of Equations 20 and 21 are simultaneously accepted, then the null hypothesis of Equation 19 would automatically be true. But if any of the null hypothesis of Equations 20 and 21 are rejected, then the null hypothesis of Equation 19 must also be rejected.

The null hypothesis of Equation 20 is equivalent to the null hypothesis that the proportion of all observations in population x_1 that are on the average greater (higher) than the common median of all the observations in populations x_1 and x_2 combined is equal to the proportion of all observations in population x_1 that are on the average smaller (lower) than the common median of the combined observations in populations x_1 and x_2. This is equivalent to testing the null hypothesis.

$$H_{01}: \ \overset{+}{\pi}_1 = \overset{-}{\pi}_1 \text{ or } \mathrm{H}_{01}: \ \overset{+}{\pi}_1 - \overset{-}{\pi}_1 = 0 \ \textit{Versus} \ H_{11}: \overset{+}{\pi}_1 - \overset{-}{\pi}_1 \neq 0. \tag{22}$$

Similarly the null hypothesis of Equation 21 is equivalent to the null hypothesis.

$$H_{02}: \ \overset{+}{\pi}_2 = \overset{-}{\pi}_2 \text{ or } \mathrm{H}_{02}: \ \overset{+}{\pi}_2 - \overset{-}{\pi}_2 = 0 \ \textit{Versus} \ H_{12}: \overset{+}{\pi}_2 - \overset{-}{\pi}_2 \neq 0. \tag{23}$$

Under the null hypothesis of equation 22, the test statistic

$$\chi_1^2 = \frac{W_1^2}{\mathrm{Var}(W_1)} = \frac{\left(f_1^+ - f_1^-\right)^2}{n_1\left(\hat{\overset{+}{\pi}}_1 + \hat{\overset{-}{\pi}}_1 - \left(\hat{\overset{+}{\pi}}_1 - \hat{\overset{-}{\pi}}_1\right)^2\right)} \tag{24}$$

under H_{01} has approximately the chi-square distribution with 1 degree of freedom. Similarly, under the null hypothesis of Equation 23 the test statistic

$$\chi_2^2 = \frac{W_2^2}{\mathrm{Var}(W_2)} = \frac{\left(f_2^+ - f_2^-\right)^2}{n_2\left(\hat{\overset{+}{\pi}}_2 + \hat{\overset{-}{\pi}}_2 - \left(\hat{\overset{+}{\pi}}_2 - \hat{\overset{-}{\pi}}_2\right)^2\right)} \tag{25}$$

has approximately the chi-square distribution with 1 degree of freedom.

The null hypothesis of Equations 22 and 23 are each rejected if the calculated chi-square values are at least equal to the tabulated or critical chi-square value with 1 degree of freedom for a specified α level; otherwise the null hypothesis is accepted.

Finally, the proposed method may be easily modified and used to test a hypothesis concerning appropriately chosen measures of central tendency for two populations when $c = \frac{\alpha}{\sqrt{3}} = 1$, and $d = \alpha = 0$.

Illustrative Example

Suppose Gender B students on the average earned one grade lower than their Gender A colleagues. On the basis of this finding, the instructor required Gender B students to mandatorily attend tutorials. The question arose whether the instructor was justified in this policy. Data were collected on a random sample of Gender A and Gender B students.

Gender A Students	B^+	A	A^-	A	C^+	A^-	A^+	A^+	C^+	A^+	C	A^+	A^-	A^+					
Gender B Students	B^+	F	F	B	A^-	D	B^+	C^+	B	B^-	A^+	B	F	E	C^+	A^+	C^-	E	C^-

First, list the Gender A students' letter grades unchanged, here designated as x_{i1}, and then list the Gender B students' grades after increasing each of them by one grade level, here designated as x_{j2}; the resulting grades are then pooled together and ranked from the highest, assigned the rank 1, to the lowest, assigned the rank 33. Tied grades are as usual assigned their mean ranks. The common median grade of the pooled sample is found to be a B^+.

Equations 1 and 5 are now applied to the listed data. The values of u_{i1}, u_{j2} and the corresponding ranks are presented in Table 1. From the values of u_{i1} in column 6 of Table 1 it is shown that

$$f_1^+ = 10, \quad f_1^0 = 1 \text{ and } f_1^- = 3 \text{ so that}$$

$$\hat{\pi}_1^+ = \frac{10}{14} = 0.714, \ \hat{\pi}_1^0 = \frac{1}{14} = 0.071 \text{ and } \hat{\pi}_1^- = \frac{3}{14} = 0.214$$

Also from Equation 11, it is found that the estimated variance of W_1 is

Table 1: Values of u_{i1}, u_{j2} and other Statistics for the illustrative Example

Gender A Grade x_{i1}	Gender B Grade x_{j2}	Adjusted Gender B Grade $x_{j2'}$	Rank of Ranking		u_{i1}	u_{i2}
			x_{i1} in the combined ranking (n_{i1})	x_{j2} in the combined ranking (n_{i1})	(Eqn1)	(Eqn5)
B^+	B^+	A^-	17.5	13	0	1
A	F	E	9	32	1	-1
A^-	F	E	13	32	1	-1
A	B	B^+	9	17.5	1	1
C^+	A^-	A	23.5	9	-1	1
A^-	D	C^-	13	28	1	-1
A^+	B^+	A^-	4	13	1	1
A^+	C^+	B^-	4	21.5	1	0
C^+	B	B^+	23.5	17.5	-1	1
A^+	B^-	B	4	20	1	1
C	A^+	A^+	26	4	-1	1
A^+	B	B^+	4	17.5	1	1
A^-	F	E	13	32	1	-1
A^+	E	D	4	29.5	1	-1
	C^+	B^-		21.5		0
	A^+	A^+		4		1
	C^-	C		26		-1
	E	D		29.5		-1
	C^-	C		26		-1
Total			**167.5**	**393.5**		

$$\begin{aligned} \text{Var}(W_1) &= 14((0.714 + 0.214) - (0.714 - 0.214)^2) \\ &= 14(0.928 - 0.25) = 14(0.678) = 9.492 \ . \end{aligned}$$

To test the null hypothesis of Equation 22 that Gender A s have the same median grades as the entire class (both genders combined), Equation 24 shows that

$$\chi_1^2 = \frac{(7)^2}{9.492} = \frac{49}{9.492} = 5.162 \quad (\text{p value} = 0.0231),$$

which with 1 degree of freedom is statistically significant, leading to a rejection of the null hypothesis. From the values of u_{j2} in column 7 of Table 1

$$f_2^+ = 9, \quad f_2^0 = 2 \text{ and } f_2^- = 8 \text{ so that}$$
$$\hat{\pi}_2^+ = \frac{9}{19} = 0.474, \ \hat{\pi}_2^0 = \frac{2}{19} = 0.105 \text{ and } \hat{\pi}_2^- = \frac{8}{19} = 0.421.$$

From Equation 8, $W_2 = f_2^+ - f_2^- = 9 - 8 = 1$. The estimated variance of W_2 is from Equation 15

$$\begin{aligned} \text{Var}(W_2) &= 19((0.474 + 0.421) - (0.474 - 0.421)^2) \\ &= 19(0.895 - 0.003) = 19(0.892) = 16.948 . \end{aligned}$$

The test statistic for the null hypothesis of Equation 23 that Gender B students have the same median score as the overall class in the course is from Equation 25.

$$\chi_2^2 = \frac{(1)^2}{16.948} = 0.059 \quad (\text{p value} = 0.8081)$$

which with 1 degree of freedom is not statistically significant, leading to an acceptance of the null hypothesis of Equation 23. Because the null hypothesis of Equations 22 and 23 are not both accepted, the null hypothesis of Equation 19 cannot be accepted. It can therefore be concluded on the basis of these tests that the hypothesized relationship between Gender A and Gender B performances in the course may not be valid.

Note that the median grade in the course for Gender A students is about an A whereas the unadjusted or original median grade for Gender B students is about a C^+ in the fabricated example. Hence, if the hypothesized relative relationship between Gender A and Gender B student grades were to hold then the original median grade for Gender B students would be expected to be about an A^-, which the adjusted Gender B median grade does not attain.

If the Wilcoxon Mann Whitney test had been used to analyze the data, it could be shown with $R_1 = 167.5$ (see Table 1) that

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (14)(19) + \frac{14(15)}{2} - 167.5$$

= 266 + 105 – 167.5 = 203.5 with mean $\mathrm{E}(U) = \frac{n_1 n_2}{2} = \frac{(14)(19)}{2} = 133$

and

$$\mathrm{Var}(u) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{(14)(19)(34)}{12} = 753.667 .$$

Thus, $Se(u) = \sqrt{753.667} = 27.453$.

The test statistic for the null hypothesis of Equation 19 for the equality of the two population medians is

$$Z = \frac{u - E(u)}{Se(u)} = \frac{203.5 - 133}{27.453} = \frac{70.5}{27.453} = 2.568 \qquad \text{(p value = 0.0102)}$$

which is statistically significant with nominal alpha set to 0.05.

The Wilcoxon Mann Whitney test, like the proposed test statistic, retained as tenable the null hypothesis of Equation 19. However, use of the usual median test with the data yielded a Chi-squared value of 0.24, which was not statistically significant, and led to an acceptance of the null hypothesis of equal population medians. This is probably due to the occasional inability of the usual median test to adequately provide for the presence of ties between the sampled populations, which may lead to an acceptance of a false null hypothesis.

From the values of u_{i1} in column 6 of the Table 2 it is shown that

$$f_1^{+} = 10, \quad f_1^{0} = 1 \text{ and } f_1^{-} = 3 \text{ so that}$$

$$\hat{\pi}_1^{+} = \frac{10}{14} = 0.714 , \ \hat{\pi}_1^{0} = \frac{1}{14} = 0.071 \text{ and } \hat{\pi}_1^{-} = \frac{3}{14} = 0.214 .$$

From Equation 4 $W_1 = f_1^{+} - f_1^{-} = 10 - 3 = 7$. From Equation 11 it is shown that the estimated variance of W_1 is

$$\mathrm{Var}(W_1) = 14((0.714 + 0.214) - (0.714 - 0.214)^2)$$
$$= 14(0.928 - 0.25) = 14(0.678) = 9.492 .$$

Table 2: Values of u_{i1}, u_{j2} and other Statistics for simulated data

Gender A Grade x_{i1}	Gender B Grade x_{j2}	Adjusted Gender B Grade $x_{j2'}$	Rank of Ranking		u_{i1}	u_{i2}
			x_{i1} in the combined ranking (n_{i1})	x_{j2} in the combined ranking (n_{i1})	(Eqn1)	(Eqn5)
B+	C-	B+	15	26.5	1	-1
A	C+	E	8.5	22.5	1	0
A+	E	A-	4	29.5	1	-1
A-	B+	C	11.5	15	1	1
C	B+	A+	25	15	-1	1
A+	D	B-	4	28	1	-1
A+	F	B+	4	32	1	-1
A-	A-	A+	11.5	11.5	0	1
A	C+	A	8.5	22.5	1	0
A-	B	D	11.5	18	1	1
C+	B-	B-	22.5	20	-1	1
A+	E	A+	4	29.5	1	-1
C+	B	C	22.5	18	-1	1
A+	F	B+	4	32	1	-1
	B	B		18		1
	A+	D		4		1
	A+	E		4		1
	C-	C-		26.5		-1
	F	E		32		-1
Total			**156.5**	**404.5**		

To test the null hypothesis of Equation 22 that Gender A students have the same median grades as the entire class (both genders combined) it can be shown from Equation 24 that

$$\chi_1^2 = \frac{(7)^2}{9.492} = \frac{49}{9.492} = 5.162 \quad (\text{p value} = 0.0231)\,,$$

which with 1 degree of freedom is statistically significant, leading to a rejection of the null hypothesis.

Furthermore, from the values of u_{j2} in column 7 of Table 2 it is shown that

$f_2^+ = 9,\ f_2^0 = 2$ and $f_2^- = 8$ so that

$\overset{+}{\pi}_2 = \frac{9}{19} = 0.474$, $\overset{0}{\hat{\pi}}_2 = \frac{2}{19} = 0.105$ and $\overset{-}{\hat{\pi}}_2 = \frac{8}{19} = 0.421$.

From Equation 8 $W_2 = f_2^+ - f_2^- = 9 - 8 = 1$. The estimated variance of W_2 from Equation 15 is

$$\begin{aligned} \text{Var}(W_2) &= 19((0.474 + 0.421) - (0.474 - 0.421)^2) \\ &= 19(0.895 - 0.003) = 19(0.892) = 16.948 . \end{aligned}$$

The test statistic for the null hypothesis of Equation 23 that Gender B students have the same median score as the overall class in the course is from Equation 25.

$$\chi_2^2 = \frac{(1)^2}{16.948} = 0.059 \quad \text{(p value = 0.8081)}$$

which with 1 degree of freedom is not statistically significant, leading to an acceptance of the null hypothesis of Equation 23.

Because the null hypothesis of Equations 22 and 23 are not both accepted, the null hypothesis of Equation 19 cannot be accepted. It can therefore be concluded on the basis of these tests that the hypothesized relationship between Gender A and Gender B performances in the course may not be valid.

Note that the median grade in the course for Gender A students is about an A whereas the unadjusted or original median grade for Gender B students is about a C^+. Hence if the hypothesized relative relationship between Gender A and Gender B student grades were to hold, then the original median grade for Gender B students would be expected to be about an A^-, which the adjusted Gender B median grade does not attain.

Using the Wilcoxon Mann Whitney U test to analyze the data would have the result, with $R_1 = 156.5$ (see Table 2), that

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (14)(19) + \frac{14(15)}{2} - 156.5$$

$$= 266 + 105 - 156.5 = 214.5 \text{ with mean } \mathrm{E}(U) = \frac{n_1 n_2}{2} = \frac{(14)(19)}{2} = 133$$

and variance.

$$\text{Var}(u) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{(14)(19)(34)}{12} = 753.667 .$$

Hence, $Se(u) = \sqrt{753.667} = 27.453$.

The test statistic for the null hypothesis of Equation 19 for the equality of the two population medians is

$$Z = \frac{u - E(u)}{Se(u)} = \frac{214.5 - 133}{27.453} = \frac{81.5}{27.453} = 2.969 \qquad (\text{p value} = 0.0030) ,$$

which is statistically significant, and the discussion on the previous example is repeated.

Conclusion

A non-parametric statistical method for the analysis of two sample data was presented that may be applied on measurements on as low as the ordinal scale and need not be homogeneous. The test statistic is intrinsically and structurally adjusted to provide for the possibility of any tied observations between the sampled populations and hence obviates the need to require the populations to be continuous. When the null hypothesis is rejected, it indicates which of the sampled populations may have been responsible for the rejection (a determination which the Wilcoxon Mann Whitney test cannot handle). Results from an example suggest that the test statistic may be as powerful as the Wilcoxon Mann Whitney test, and more powerful than the usual median test.

References

Afuecheta, E. O., Oyeka, I. C. A, Ebuh, G. U., & Nnanatu, C. C. (2012). Modified Median Test Intrinsically Adjusted For Ties. *Journal of Basic Physical Research, 3*, 30-34.

Ebuh, G. U. & Oyeka, I. C. A. (2012). Statistical Comparison of Eight Alternative Methods for the Analysis of Paired Sample Data with Applications. *Open Journal of Statistics (OJS), 2*(3), 328-345.

Ebuh, G. U., Oyeka, I. C. A., & Nwosu, C. R. (2012). Application of Dummy Variables Multiple Regression to Paired Samples with three Options. *Journal of Nigerian Statistical Association (JNSA) 24*, 34-44

Gibbons, J. D. (1971). *Non-parametric statistical inference*. New York: McGraw Hill.

Oyeka, I. C. A., Ebuh, G. U., Nwosu, C. R., Utazi, E. C., Ikpegbu, P. A., Obiora-Ilouno, H., & Nwankwo, C. C. (2009). A Method of Analysing Paired Data Intrinsically Adjusted for Ties. *Global Journal of Mathematics and Statistics, India*. *1*(1), 1-6

Oyeka, I. C. A. (2009). *An Introduction to applied statistical methods*, 8th edition. Enugu: Nobern Avocation Publishing Company, 457-495.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 171-183.

ISSN 1538 − 9472

Test for Intraclass Correlation Coefficient under Unequal Family Sizes

Madhusudan Bhandary
Columbus State University
Columbus, GA

Koji Fujiwara
North Dakota State University
Fargo, ND

Three tests are proposed based on F-distribution, Likelihood Ratio Test (LRT) and large sample Z-test for intraclass correlation coefficient under unequal family sizes based on a single multinormal sample. It has been found that the test based on F-distribution consistently and reliably produces results superior to those of Likelihood Ratio Test (LRT) and large sample Z-test in terms of size for various combinations of intraclass correlation coefficient values. The power of this test based on F-distribution is competitive with the power of the LRT and the power of Z-test is slightly better than the powers of F-test and LRT when k = 15, but the power of Z-test is worse in comparison with the F-test and LRT for k = 30, i.e. for large sample situation, where k = sample size. This test based on F-distribution can be used for both small sample and large sample situations. An example with real data is presented.

Keywords: Likelihood ratio test, Z-test, F-test, intraclass correlation coefficient.

Introduction

Suppose it is required to estimate the correlation coefficient between blood pressures of children on the basis of measurements taken on p children in each of n families. The p measurements on a family provide $p(p\text{-}1)$ pairs of observations, (x,y) --- x being the blood pressure of one child and y that of another. From the n families a total of $np(p\text{-}1)$ pairs are generated from which a correlation coefficient is computed in the ordinary way.

The correlation coefficient thus computed is called an intraclass correlation coefficient. It is important to have statistical inference concerning intraclass correlation, because it provides information regarding blood pressure, cholesterol, etc., in a family within some race in the world.

Madhusudan Bhandary is a Professor in the Department of Mathematics. Email him at: bhandary_madhusudan@colstate.edu. Koji Fujiwara is a graduate student in the Department of Statistics. Email him at: koji.fujiwara@ndsu.edu.

The intraclass correlation coefficient ρ has a wide variety of uses in measuring the degree of intrafamily resemblance with respect to characteristics such as blood pressure, cholesterol, weight, height, stature, lung capacity, etc.

Several authors have studied statistical inference concerning ρ based on a single multinormal sample (Scheffe, 1959; Rao, 1973; Rosner et.al, 1977, 1979,; Donner and Bull, 1983; Srivastava, 1984; Konishi, 1985; Gokhale and SenGupta, 1986; SenGupta, 1988; Velu and Rao, 1990). Donner and Bull (1983) discussed the likelihood ratio test for testing the equality of two intraclass correlation coefficients based on two independent multinormal samples under equal family sizes. Konishi and Gupta (1987) proposed a modified likelihood ratio test and derived its asymptotic null distribution. They also discussed another test procedure based on a modification of Fisher's Z-transformation following Konishi (1985).

Huang and Sinha (1993) considered an optimum invariant test for the equality of intraclass correlation coefficients under equal family sizes for more than two intraclass correlation coefficients based on independent samples from several multinormal distributions. For unequal family sizes, Young and Bhandary (1998) proposed Likelihood ratio test, large sample Z-test and large sample Z^*-test for the equality of two intraclass correlation coefficients based on two independent multinormal samples.

For several populations and unequal family sizes, Bhandary and Alam (2000) proposed Likelihood ratio test and large sample ANOVA test for the equality of several intraclass correlation coefficients based on several independent multinormal samples. Donner and Zou (2002) proposed asymptotic test for the equality of dependent intraclass correlation coefficients under unequal family sizes.

However, none of the above authors derived any test for a single sample and unequal family sizes. It is an important practical problem to consider a single sample test for intraclass correlation coefficient under unequal family sizes.

This article considers three tests for intraclass correlation coefficient based on a single multinormal sample under unequal family sizes. Conditional analysis is conducted assuming family sizes fixed though unequal. It could be of interest to examine the blood pressure or cholesterol or lung capacity, etc., among families in U.S.A. or among some other races, and therefore it is necessary to develop a single sample test for intraclass correlation coefficient under unequal family sizes.

Three tests are proposed: F-test, LRT and large sample Z-test. These three tests are compared in Section 3 using simulation technique. It has been found on the basis of simulation study that the test based on F-distribution consistently and

reliably produced results superior to those of Likelihood Ratio Test (LRT) and large sample Z-test in terms of size for various combinations of intraclass correlation coefficient values.

The power of this test based on F-distribution is competitive with the power of the LRT, and the power of Z-test is slightly better than the powers of F-test and LRT when k=15; but the power of Z-test is worse in comparison with the F-test and LRT for k=30, i.e. for large sample situation, where k = sample size.

This test based on F-distribution can be used for both small sample and large sample situations.

An example with real data is presented in Section 4.

Tests of $H_0: \rho = \rho_0$ Versus $H_1: \rho \neq \rho_0$

Likelihood Ratio Test:

Let $\underset{\sim}{X}_i = (x_{i1}, x_{i2}, ..., x_{ip_i})'$ be a $p_i x 1$ vector of observations from i^{th} family;

$$i = 1, 2, ..., k.$$

The structure of mean vector and the covariance matrix for the familial data is given by the following (Rao 1973) :

$$\underset{\sim}{\mu}_i = \mu \underset{\sim}{1}_i \text{ and } \underset{p_i x p_i}{\Sigma_i} = \sigma^2 \begin{pmatrix} 1 & \rho & ... & \rho \\ \rho & 1 & ... & \rho \\ ... & ... & ... & ... \\ \rho & \rho & ... & 1 \end{pmatrix} \tag{1}$$

where $\underset{\sim}{1}_i$ is a $p_i x 1$ vector of 1's, $\mu(-\infty < \mu < \infty)$ is the common mean and $\sigma^2(\sigma^2 > 0)$ is the common variance of members of the family and ρ, which is called the intraclass correlation coefficient, is the coefficient of correlation among the members of the family and $\max_{1 \le i \le k} \left(-\frac{1}{p_i - 1} \right) \le \rho \le 1.$

It is assumed that $\underset{\sim}{x}_i \sim N_{p_i}(\underset{\sim}{\mu}_i, \Sigma_i); i = 1, ..., k$, where N_{p_i} represents p_i -variate normal distribution and $\underset{\sim}{\mu}_i, \Sigma_i$'s are defined in (1).

Let $\underset{\sim}{u}_i = (u_{i1}, u_{i2}, ..., u_{ip_i})' = Q \underset{\sim}{x}_i$ (2)

where Q is an orthogonal matrix.

Under the orthogonal transformation (2), it can be seen that $\underset{\sim}{u}_i \sim N_{p_i}(\underset{\sim}{\mu}_i^*, \Sigma_i^*); i = 1, ..., k$

where $\underset{\sim}{\mu}_i^* = (\mu, 0, ..., 0)'$ and $\Sigma_i^* = \sigma^2 \begin{pmatrix} \eta_i & 0 & ... & 0 \\ 0 & 1-\rho & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1-\rho \end{pmatrix}$

and $\eta_i = p_i^{-1}\{1 + (p_i - 1)\rho\}$

The transformation used on the data from $\underset{\sim}{x}$ to $\underset{\sim}{u}$ above is independent of ρ. One can use Helmert's orthogonal transformation.

Srivastava (1984) gives the estimator of ρ and σ^2 under unequal family sizes which are good substitutes for the maximum likelihood estimators and are given by the following:

$$\hat{\rho} = 1 - \frac{\hat{\gamma}^2}{\hat{\sigma}^2}$$

$$\hat{\sigma}^2 = (k-1)^{-1} \sum_{i=1}^{k} (u_{i1} - \hat{\mu})^2 + k^{-1} \hat{\gamma}^2 \left(\sum_{i=1}^{k} a_i\right)$$

$$\hat{\gamma}^2 = \frac{\sum_{i=1}^{k} \sum_{r=2}^{p_i} u_{ir}^2}{\sum_{i=1}^{k} (p_i - 1)} \qquad (3)$$

$$\hat{\mu} = k^{-1} \sum_{i=1}^{k} u_{i1}$$

and $a_i = 1 - p_i^{-1}$

Now, consider a random sample of *k* families from a population.

Let $\underset{\sim}{x}_i = (x_{i1}, x_{i2}, ..., x_{ip_i})'$ be a $p_i x1$ vector of observations from i^{th} family; $i = 1, 2, \ldots, K$

and $\underset{\sim}{x}_i \sim N_{p_i}(\underset{\sim}{\mu}_i, \Sigma_i)$, where $\underset{\sim}{\mu}_i = \mu \underset{\sim}{1}_i, \Sigma_i = \sigma^2 \begin{pmatrix} 1 & \rho & ... & \rho \\ \rho & 1 & ... & \rho \\ ... & ... & ... & ... \\ \rho & \rho & ... & 1 \end{pmatrix}$ (4)

and $\max_{1 \le i \le k_1} \left(-\frac{1}{p_i - 1} \right) \le \rho \le 1$.

Using orthogonal transformation, the data vector can be transformed from $\underset{\sim}{x}_i$ to $\underset{\sim}{u}_i$ as follows:

$$\underset{\sim}{u}_i = (u_{i1}, u_{i2}, ..., u_{ip_i})' \sim N_{p_i}(\underset{\sim}{\mu}_i^*, \Sigma_i^*); i = 1, ..., k$$

where, $\underset{\sim}{\mu}_i^* = (\mu, 0, ..., 0)', \Sigma_i^* = \sigma^2 \begin{pmatrix} \eta_i & 0 & ... & 0 \\ 0 & 1-\rho & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1-\rho \end{pmatrix}$

$$\eta_i = p_i^{-1} \{1 + (p_i - 1)\rho\} \quad (5)$$

The transformation used on the data above from $\underset{\sim}{x}$ to $\underset{\sim}{u}$ is independent of ρ.

Under the above setup, likelihood ratio test statistic for testing $H_0 : \rho = \rho_0$ Vs. $H_1 : \rho \neq \rho_0$ is given by the following:

$$-2 \log \Lambda = \sum_{i=1}^{k} \log \left[p_i^{-1} \{1 + (p_i - 1)\rho_0\} \right] + \sum_{i=1}^{k} (p_i - 1) \log(1 - \rho_0)$$

$$+\frac{1}{\hat{\sigma}^2}\left[\sum_{i=1}^{k}\left\{p_i(u_{i1}-\hat{\mu})^2/\left[1+(p_i-1)\rho_0\right]\right\}+\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2/(1-\rho_0)\right]$$
$$-\sum_{i=1}^{k}\log\left[p_i^{-1}\left\{1+(p_i-1)\hat{\rho}\right\}\right]-\sum_{i=1}^{k}(p_i-1)\log(1-\hat{\rho})$$
$$-\frac{1}{\hat{\sigma}^2}\left[\sum_{i=1}^{k}\left\{p_i(u_{i1}-\hat{\mu})^2/\left[1+(p_i-1)\hat{\rho}\right]\right\}+\sum_{i=1}^{k}\sum_{r=2}^{p_i}u_{ir}^2/(1-\hat{\rho})\right] \quad (6)$$

where, Λ = likelihood ratio test statistic,

$\hat{\rho}$ = estimate of intraclass correlation coefficient under H_1,
$\hat{\sigma}^2$ = estimate of σ^2

and $\hat{\mu}$ is the estimate of mean.

The estimators $\hat{\rho},\hat{\sigma}^2$ and $\hat{\mu}$ can be obtained from Srivastava's estimator given by (3).

It is well-known from asymptotic theory that $-2\log\Lambda$ has an asymptotic chi-square distribution with 1 degree of freedom.

Large Sample Z-test:

A large sample Z-test is proposed as follows :

$$Z=\frac{\hat{\rho}-\rho_0}{\sqrt{\frac{Var}{k}}} \quad (7)$$

where, $\hat{\rho}$ = the estimator of ρ from the sample using Srivastava (1984) and Var under H_0 (using Srivastava and Katapa (1986)) is as follows :

$$\text{Var} = 2(1-\rho_0)^2\left\{(\overline{p}-1)^{-1}+c^2-2(1-\rho_0)(\overline{p}-1)^{-1}k^{-1}(\sum_{i=1}^{k}a_i)\right\} \quad (8)$$

where, k = number of families in the sample

$$\overline{p} = k^{-1}\sum_{i=1}^{k} p_i$$

$$c^2 = 1 - 2(1-\rho_0)^2 k^{-1}\sum_{i=1}^{k} a_i + (1-\rho_0)^2\left[k^{-1}\sum_{i=1}^{k} a_i + (\overline{p}-1)^{-1}\overline{a}^2\right]$$

$$\overline{a} = k^{-1}\sum_{i=1}^{k} a_i$$

and $a_i = 1 - p_i^{-1}$

It is obvious (using Srivastava and Katapa (1986)) to see that under H_0, the test statistic Z given by (7) has an asymptotic $N(0,1)$ distribution. Because the alternative hypothesis is $H_1 : \rho \neq \rho_0$, the above Z-test is a two sided test.

F-test:

Using (5), $u_{i1} \sim N_1(\mu, \sigma^2\eta_i)$

where $\eta_i = p_i^{-1}\{1 + (p_i - 1)\rho\}$

Hence $\sum_{i=1}^{k} \frac{(u_{i1} - \hat{\mu})^2}{\sigma^2\eta_i} \sim \chi^2_{k-1}$

Also from (5),

$$\frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i} u_{ir}^2}{\sigma^2(1-\rho)} \sim \chi^2_{\sum_{i=1}^{k}(p_i - 1)}$$

and $\sum_{i=1}^{k} \frac{(u_{i1} - \hat{\mu})^2}{\sigma^2\eta_i}$ and $\frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i} u_{ir}^2}{\sigma^2(1-\rho)}$ are independent.

χ^2_n denotes Chi-square distribution with n degrees of freedom.

Hence, an F-test is proposed as:

$$F = \frac{\sum_{i=1}^{k} \frac{(u_{i1} - \hat{\mu})^2}{p_i^{-1}\{1+(p_i-1)\rho_0\}} \Big/ (k-1)}{\frac{\sum_{i=1}^{k}\sum_{r=2}^{p_i} u_{ir}^2}{(1-\rho_0)} \Big/ (\sum_{i=1}^{k}(p_i-1))} \tag{9}$$

$$\text{Under } H_0, \; F \sim F_{k-1, \sum_{i=1}^{k}(p_i-1)} \tag{10}$$

The performance of the three tests given by (6), (7) and (9) is discussed in terms of size and power in the next section using simulated data.

Simulation Results:

Multivariate normal random vectors were generated using R program in order to evaluate the power of the F statistic as compared to the LRT statistic and Z-statistic. Five and thirty vectors of family data were created for the population. The family size distribution was truncated to maintain the family size at a minimum of 2 siblings and a maximum of 15 siblings. The previous research in simulating family sizes (Rosner et al. (1977), Srivastava and Keen (1988)) determined the parameter setting for FORTRAN IMSL negative binomial subroutine with a mean = 2.86 and a success probability = 0.483.

Here, a mean = 2.86 and a theta = 41.2552 were set.

All parameters were set the same except the value of ρ which took on all combinations possible over the range of values from 0.1 to 0.9 at increments of 0.1.

The R program produced estimates of ρ along with F statistic, the LRT statistic and the Z- statistic 3,000 times for each particular population parameter ρ.

The frequency of rejection of each test statistic at α =0.05 was noted and the proportion of rejections are calculated for various combination of ρ.

The sizes for the LRT statistic, F statistic and Z statistic for various combination of ρ were also calculated.

On the basis of this study, it was found that the test based on F-distribution consistently and reliably produced results superior to those of Likelihood Ratio Test (LRT) and large sample Z-test in terms of size for various combinations of intraclass correlation coefficient values.

The power of this test based on F-distribution is competitive with the power of the LRT and the power of Z-test is slightly better than the powers of F-test and LRT when k=15 but the power of Z-test is worse in comparison with the F-test and LRT for k=30 i.e. for large sample situation.

Hence, overall recommendation should be to use F-test given by (9).

The LRT has a large bias under H_0 and this bias is probably due to the fact that $-2\log\Lambda$ may take some negative values (which it should not take) and those negative values are deleted for the calculation of size and power. But, for the F-test or Z-test, this problem does not arise.

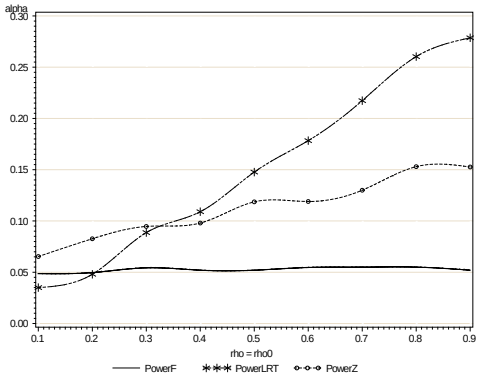


Figure 1. Alpha Levels (k = 15, alpha = 0.05)

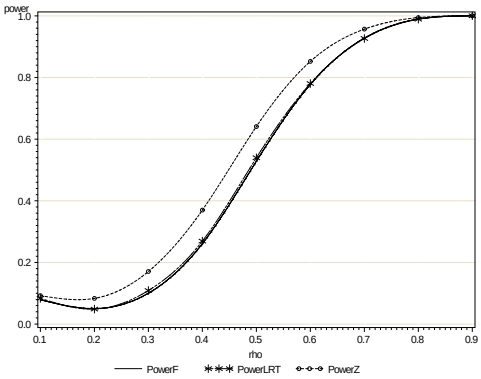


Figure 3. Power (alpha = 0.05, k = 15, rho0 = 0.2)

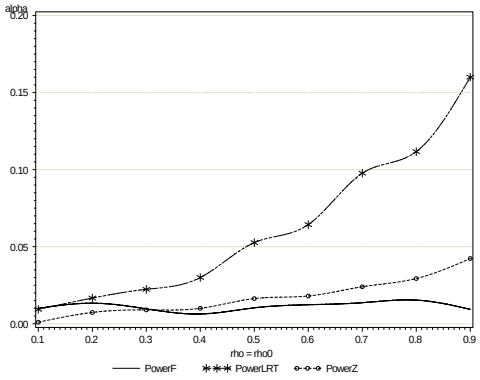


Figure 2. Alpha Levels (k = 30, alpha = 0.01)

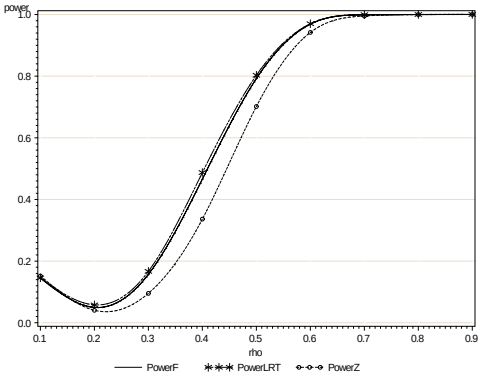


Figure 4. Power (alpha = 0.05, k = 30, rho0 = 0.2)

The test based on F-distribution can be used for both small sample and large sample situations.

Hence, the F test is strongly recommended for use in practice.

Example with Real Life Data

The three tests are compared using real life data collected from Srivastava and Katapa (1986). Table 1 gives the values of pattern intensity on soles of feet in fourteen families, where values for daughters and sons are put together. In the Table below, for example, for family #11, the children (both sons and daughters) have feet sizes 5,3,4,4 respectively.

Table 1. Pattern intensity values on soles of feet for 14 families.

Sample	Family	Siblings
A	12	2, 4
A	10	4, 5, 4
A	9	5, 6
A	1	2, 2
A	4	2, 2, 2, 2, 2
A	5	6, 6
A	8	2, 4, 7, 4, 4, 7, 8
A	3	2, 2, 2
A	6	4, 3, 3
A	14	2, 2, 2
A	7	2, 2, 3, 6, 3, 5, 4
A	2	2, 3
A	11	5, 3, 4, 4
A	13	4, 3, 3, 3

The data on the children from Table 1 was used to analyze the case of testing intraclass correlation coefficient. The above data were collected for the purpose of having inference on familial correlation. Srivastava and Katapa (1986) used the above data to estimate the intraclass correlation coefficient for unequal family sizes.

First, the data is transformed by multiplying each observation vector by Helmert's orthogonal matrix Q

where, $$\underset{p_i x p_i}{Q} = \begin{bmatrix} \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \ldots & \frac{1}{\sqrt{p_i}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \ldots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \ldots & -\frac{(p_i-1)}{\sqrt{p_i(p_i-1)}} \end{bmatrix}$$

This results in transformed vectors $\underset{\sim}{u_i}$ for $i = 1, 2, ..., k$. Here, k =14.

Srivastava's formula given by (3) is used to compute intraclass correlation coefficient and variance. The computed values of intraclass correlation coefficient and variance are $\hat{\rho} = 0.8118$ and $\hat{\sigma}^2 = 8.8578$. Because $\hat{\rho} = 0.8118$ is estimated from the above sample, it is necessary to know whether the intraclass correlation coefficient ρ in the population from which the sample came is close to 0.8, and therefore necessary to test $H_0 : \rho = 0.8$ Vs. $H_1 : \rho \neq 0.8$.

Formulae (6) and (7) and (9) are used to obtain the values of the test statistics for testing $H_0 : \rho = 0.8$ Vs. $H_1 : \rho \neq 0.8$. The computed values of the LRT statistic, Z statistic and F statistic obtained from formula (6), (7) and (9) respectively are as follows:

LRT statistic = 331.31, Z statistic = 0.10642 and F statistic = 0.7316

The critical values at α = 0.05 and 0.10 for the tests are as follows:

$$LRT_{0.05} = 3.8415; Z_{0.025} = 1.96; F_{0.025} = 2.8506; F_{0.975} = 0.33037;$$
$$LRT_{0.10} = 2.7055; Z_{0.05} = 1.645; F_{0.05} = 2.3973; F_{0.95} = 0.39763.$$

Hence, the null hypothesis is accepted by F- test and Z- test at 5% and 10% levels, whereas it is rejected by LRT at 5% and 10% levels.

If the F-test is used it provides information that the population intraclass correlation coefficient is 0.8 which should be true intuitively. But, LRT gives information that the population intraclass correlation coefficient is not 0.8. Therefore, it is important to choose the right test to produce the correct information. It was found before that LRT is biased in terms of size; therefore the recommendation is to use F-test given by (9) in real practice.

References

Bhandary, M., & Alam, M. K. (2000). Test for the equality of intraclass correlation coefficients under unequal family sizes for several populations. *Communications in Statistics-Theory and Methods, 29*(4), 755-768.

Donner, A., & Bull, S. (1983). Inferences concerning a common intraclass correlation coefficient. *Biometrics*, *39*, 771-775.

Donner, A. , & Zou, G. (2002). Testing the equality of dependent intraclass correlation coefficients. *The Statistician, 51*(3), 367-379.

Gokhale, D. V., & SenGupta, A. (1986). Optimal tests for the correlation coefficient in a symmetric multivariate normal population. *J. Statist. Plann Inference*, *14*, 263-268.

Huang, W., & Sinha, B. K. (1993). On optimum invariant tests of equality of intraclass correlation coefficients. *Annals of the Institute of Statistical Mathematics*, *45*(3), 579-597.

Konishi, S. (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics*, *37*, 87-94.

Konishi, S. , & Gupta, A. K. (1987). Testing the equality of several intraclass correlation coefficients. *J. Statist. Plann. Inference, 21*, 93-105.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.

Rosner, B., Donner, A., & Hennekens, C.H. (1977). Estimation of intraclass correlation from familial data. *Applied Statistics, 26*, 179-187.

Rosner, B., Donner, A., & Hennekens, C.H. (1979). Significance testing of interclass correlations from familial data. *Biometrics, 35*, 461-471.

Scheffe, H. (1959). *The Analysis of Variance.* Wiley, New York.

SenGupta, A. (1988). On loss of power under additional information – an example. *Scand. J. Statist., 15*, 25-31.

Srivastava, M. S. (1984). Estimation of interclass correlations in familial data. *Biometrika, 71,* 177-185.

Srivastava, M.S. , & Katapa, R. S. (1986). Comparison of estimators of interclass and intraclass correlations from familial data. *Canadian Journal of Statistics, 14,* 29-42.

Srivastava, M. S. , & Keen, K. J. (1988). Estimation of the interclass correlation coefficient. *Biometrika, 75,* 731-739.

Velu, R. , & Rao, M. B. (1990). Estimation of parent-offspring correlation. *Biometrika, 77*(3), 557-562.

Young, D., & Bhandary, M. (1998). Test for the equality of intraclass correlation coefficients under unequal family sizes. *Biometrics, 54*(4), 1363-1373.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 184-190.

ISSN 1538 − 9472

Variables Sampling Plan For Correlated Data

J. R. Singh
Vikram University
Ujjain, India

R. Sankle
Vikram University
Ujjain, India

M. Ahmad Khanday
Vikram University
Ujjain, India

The sampling plan for the mean for correlated data is studied. The Operating Characteristic (OC) of the variable sampling plan for mean for correlated data are calculated and compared with the OC of known σ case.

Keywords: Variable sampling plan, correlation coefficient, operating characteristic function

Introduction

Quality control methods are commonly used to determine acceptability of a product with regard to its usefulness at the time it is put into service. It is essential from the consumer point of view, however, that a product return its usefulness for a certain length of time. Acceptance sampling is the testing or the inspection of selected items from a given lot followed by acceptance or rejection of that lot on the basis of the results of the test and its indicator of the lot's quality. It is assumed that a lot's quality is determined by the proportion of defective items in the lot. Further, attention will be restricted to those types of defects that are determined by one sided specification limits. For the purpose of exactness, upper specification limits will usually be discussed, i.e. an item will said to be defective if its measured characteristics are greater than some specified value *U*. Variable plans, however, require that the characteristic of interest be continuous variable. The characteristic is measured and its actual value is recorded. In variable sampling plans an underlying process distribution form is assumed. Then the proportion defective in the lot can be estimated by estimating the parameters of that distribution. The variable model thus requires more restrictive assumptions on the manufacturing process. If these assumptions can be justified there would be a substantial saving in sample size corresponding to a given sampling list.

The authors are in the School of Studies in Statistics. Email them, respectively, at jrsinghstat@gmail.com, rajshreesankle@gmail.com, and manzoorstat@gmail.com (corresponding author).

When the standard deviation of the lot quality is known, the criteria for acceptance and the associated mathematical computations get simplified. But, we should examine in each case whether treating the lot standard deviation (σ) as known and giving it a particular value are justified. When products are manufactured by automatic machinery whose inherent variation is known and tested, an example is provided where the lot standard deviation is known. When it is assumed that the lot standard deviation is known, and given a particular value σ, it must be remembered that σ is a constant in calcu1ations and discussions. Also, the previous assumption that the lots are formed in such a way as to ensure homogeneity within lots holds good here also; and we assume that the directly measurable quality X follows the normal law of pattern of variation in the lot; these assumptions must be examined and reviewed from time to time when variables plans with known σ are in use.

Hapuaxachi and Macephexsan (1992) studied the effect of serial correlation on acceptance sampling plans by variables by comparing Operating Characteristic (OC) curves, sample size and producers risk, α with that of the independent case when the process standard deviation (σ) is known. When σ is unknown and for large n, sampling plans can be constructed using central limit theorem. Several works have studied the effect of correlated data (see Kaiyang & Hancock, 1990; Seal, 1959; and Qiu et al., 2010). This study examines the sampling plan for mean for correlated data. The OC function of the variable sampling plan for mean for correlated data are calculated and compared with the OC function of known σ case.

Model Description And OC Function For Correlated Data

For a single sampling plan, with known-sigma, the procedure of selection of sample is as in the other single sampling plans. The n units in the sample are measured, and the values x_1, x_2, x_3,..., x_n are obtained. The mean $\overline{x}$ is calculated. Since the standard deviation sigma (σ) of the lot is known, σ is used.

Suppose that observations x_1, x_2, x_3, ..., x_n have a multivariate normal distribution with $E(x_i) = \mu$ and $Var(x_i) = \sigma^2$ and ρ as the common correlation coefficient between any x_i and x_j, $i \neq j$. Then

$$E(\overline{x}) = \mu \tag{1}$$

and

$$\begin{aligned} Var(\bar{x}) &= \frac{\sigma^2}{n}\left[1+(n-1)\rho\right] \\ &= \frac{\sigma^2}{n}T^2 \end{aligned} \tag{2}$$

where

$$T^2 = \left[1+(n-1)\rho\right]. \tag{3}$$

In connection with a single sampling variable plan, when data are correlated, the following symbols will be used,

L = Lower specification limit,
U = Upper specification limit,
k = Acceptance parameter,
$\bar{x}$ = Sample mean of correlated data,
ρ = Correlation Coefficient.

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \tag{4}$$

where $z \sim N(0,1)$.

The OC function of single sampling plan can now be calculated. The acceptance criterion for correlated data mean plan is, for upper specification limit U, accept the lot if

$$\bar{x} + \frac{k\sigma T}{\sqrt{n}} \le U \tag{5}$$

reject the lot otherwise.

The values of n and k are determined for a given set of values of the producer risk, α and consumer risk, β, AQL and LTPD, by formulae

$$n = \left[\frac{\left(K_\alpha + K_\beta\right)}{\left(K_{p_1} - K_{p_2}\right)}\right]^2 \tag{6}$$

$$k = \left[\frac{K_\alpha K_{p_2} + K_\beta K_{p_1}}{K_\alpha + K_\beta} \right] \tag{7}$$

If p is the proportion defective in the lot

$$\frac{U - \mu}{\sigma} = K_p, \tag{8}$$

where $p = \int_{K_p}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$.

The expression for probability of acceptance (OC) function of the plan in normal case is

$$L(p) = Prob\left[\bar{x} + \sqrt{MSE(\bar{x})} \le U = \mu + K_p \sigma \frac{T}{\sqrt{n}} \right] \tag{9}$$

where

$$T^2 = [1 + (n - 1)\rho].$$

Following Schilling (1982) the OC function for correlated data works out to be as

$$L(p) = \Phi\left[\frac{\sqrt{n}}{T}\left(K_p - k\right) \right] \tag{10}$$

where

$$\Phi(t) = \int_{-\infty}^{t} \phi(z)\, dz$$

The usual single sampling plan for known σ is

$$L(p) = \Phi\left[\sqrt{n}\left(K_p - k\right)\right] \qquad (11)$$

Numerical Illustration and Result

For illustration, consider an example of producers and consumers oriented single sampling plan $p_1 = 0.01$, $\alpha = 0.05$, $p_2 = 0.08$, and $\beta = 0.10$. The values of n and k have been determined from equation (6) and (7) and are 10 and 18.09, respectively. The values of OC function for the above plan have been calculated for correlated data as well as for known standard deviation by using equation (10) and (11). These values of OC function for different values of correlation and usual known standard deviation case are presented in Table 1 and are plotted in Figure 1.

Table 1. Values of OC for different values of ρ

P	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
0.005	0.9926	0.8944	0.8056	0.7867	0.7574	0.7354
0.010	0.9500	0.8006	0.7196	0.7042	0.6812	0.6645
0.020	0.7820	0.6553	0.6085	0.6003	0.5882	0.5797
0.030	0.5908	0.5469	0.5323	0.5298	0.5262	0.5236
0.040	0.4271	0.4625	0.4741	0.4761	0.4790	0.4811
0.050	0.3016	0.3949	0.4271	0.4327	0.4408	0.4466
0.060	0.2101	0.3396	0.3878	0.3963	0.4087	0.4176
0.080	0.1000	0.2555	0.3252	0.3380	0.3568	0.3704
0.100	0.0471	0.1953	0.2770	0.2925	0.3159	0.3328
0.120	0.0221	0.1510	0.2384	0.2558	0.2823	0.3017
0.140	0.0104	0.1177	0.2067	0.2253	0.2539	0.2752
0.160	0.0049	0.0923	0.1803	0.1995	0.2296	0.2522
0.180	0.0023	0.0728	0.1579	0.1774	0.2083	0.2319
0.200	0.0011	0.0576	0.1388	0.1582	0.1896	0.2139

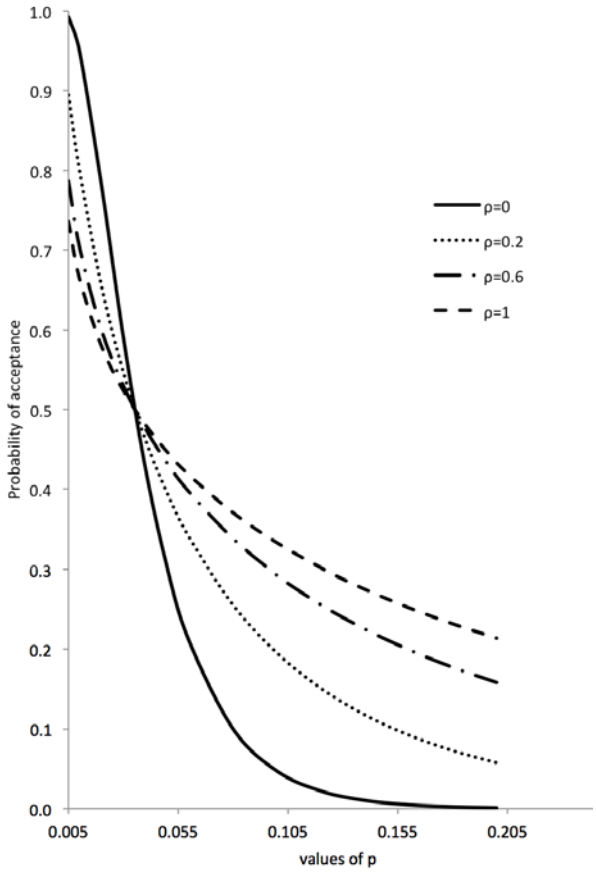


Figure 1. OC Curve for different values of Correlated Data (ρ)

From Figure 1 it is evident that the effect of correlation data on OC increases as ρ increases. As ρ increases, a significant effect is seen in producers risk as well as consumers risk, which is not acceptable. Hence one should maintain the correlation between the observations as low as possible, so as to protect producer as well as consumer.

References

Hapuaxachchi, K. P., & Macpherson, B. D. (1992). Autoregressive process applied to acceptance sampling by variables. *Communication in Statistics – Simulation and Computation, 21*(3), 833-848.

Qiu, P., Zou. C., & Wang, Z. (2010). Nonparametric profile monitoring by mixed effects modeling. *Technometrics, 52*(3), 265-277.

Schilling, E. G. (1982). *Acceptance Sampling in Quality Control.* New York: Marcel Dekker, Inc.

Seal, K. C. (1959). A single sampling plan for correlated variable with a single sided specification limit. *Journal of American Statistical Association., 54*(285), 248-259.

Yang, K., & Hancock, W. M. (1990). Statistical quality control for correlated samples. *International Journal of Production Research, 28*(3), 595-608.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 191-210.

ISSN 1538 − 9472

Case-Control Studies with Jointly Misclassified Exposure and Confounding Variables

Tze-San Lee
Western Illinois University
Macomb, IL

The issue of 2 × 2 × 2 case-control studies is addressed when both exposure and confounding variables are jointly misclassified. Two scenarios are considered: the classification errors of exposure and confounding variables are independent or not independent. The bias-adjusted cell probability estimates which account for the misclassification bias are presented. The effect of misclassification on the measure of crude odds ratio either unstratified or stratified by the confounder, Mantel-Haenszel summary odds ratio, the confounding component in the crude odds ratio, the first and second order multiplicative interaction are assessed through the sensitivity analysis from using the data on the asthma deaths of 5-45 aged patients in New Zealand.

Keywords: Asthma mortality, confounding, effect modification, Mantel-Haenszel summary odds ratio, multiplicative interaction.

Introduction

Misclassification is a ubiquitous problem in epidemiologic studies. A 2 × 2 case-control study with a single exposure variable being misclassified has been thoroughly studied (Fleiss et al. 2003, Chapter 17; Gustafson 2004, Chapter 5; Kleinbaum et al. 1982, Chapter 12; Rothman et al 2008, Chapter 19). In contrast, the misclassification of a confounding factor has attracted less attention, although there are some important papers on this topic (Ahlbom & Steineck 1992; Axelson 1978; Greenland 1980; Greenland & Robins 1985; Kupper 1984; Savitz & Baron 1989; Walker 1985). However, few papers address the issue when the study (or exposure) factor and the confounding factor are simultaneously misclassified. Most articles focused merely on the aspect that the confounding factor is misclassified.

Tze-San Lee is presently working as a mathematical statistician at the Centers for Disease Control and Prevention in Chamblee, GA. Email at tjl3@cdc.gov.

Although Fung & Howe (1984) considered the joint misclassification of polytomous exposure and confounding variables, they do not provide the bias-adjusted estimator for the cell probability. Tzonou et al. (1986) investigates the effect of misclassification on the summary odds ratio in case-control studies in which both exposure and confounding variables are jointly misclassified. But they merely consider the scenario that the classification errors of the exposure and confounding factors are independent. Again, no bias-adjusted estimators are provided in their paper.

The scenarios are addressed here in which the joint classification errors of the exposure and confounding factors are either independent or not independent. Below, necessary background materials are first reviewed. The misclassification probabilities are then defined. The formulas for all bias-adjusted measures of the effect caused by the joint misclassification of exposure and confounder are thus presented. A real-world data set is used as an example to illustrate how to calculate the misclassification probabilities by employing the counterfactual (or correctly classified) tables when the validation data are not available. A sensitivity analysis is then carried out for the admissible counterfactual tables.

Methodology & Background

Let D, E, and C be three dichotomous variables, in which D denotes the subject's outcome (disease) variable (=1 if present, 0 otherwise), E the subject's exposure variable (= 1 if exposed, 0 otherwise), and C (= 1 if present, 0 otherwise) the extraneous (a suspected confounding) variable. Assume that a simple random sampling scheme is used to collect the data of size n which are then cross-classified into table 1 in which E^* and C^* are imperfect classification variables for E and C.

Table 1: Observed contingency table of three dichotomous variables $D \times E \times C$

		$C^* = 1$	$C^* = 0$	Total
$D = 1$ Cases)	$E^* = 1$	n_{111}	n_{110}	n_{11+}
	$E^* = 0$	n_{101}	n_{100}	n_{10+}
	Total	$\boldsymbol{n_{1+1}}$	$\boldsymbol{n_{1+0}}$	$\boldsymbol{n_{1++}}$
$D = 0$ (Controls)	$E^* = 1$	n_{011}	n_{010}	n_{01+}
	$E^* = 0$	n_{001}	n_{000}	n_{00+}
	Total	$\mathbf{n_{0+1}}$	$\mathbf{n_{0+0}}$	$\mathbf{n_{0++}}$

The crude cell probability estimators are given by: for fixed i

$$\hat{p}_{ijk} = n_{ijk} / n_{(i)} \tag{1}$$

where $n_{(i)} \equiv n_{i++} = n_{i11} + n_{i10} + n_{i01} + n_{i00}$, the "+" sign in the subscript represents summation in the usual way.

It is assumed that Eq. 1 follows a multinomial distribution with parameters $n_{(i)}$ and p_{ijk} . For fixed i, the variance-covariance matrix of $\{ \hat{p}_{ijk} \}$ is given by

$$\Sigma_{(i)} = [\sigma_{jk(i)}]_{j,k=1}^{4} = n_{(i)} \cdot \begin{bmatrix} p_{i11} \cdot q_{i11} & -p_{i11} \cdot p_{i10} & -p_{i11} \cdot p_{i01} & -p_{i11} \cdot p_{i00} \\ \sigma_{12(i)} & p_{i10} \cdot q_{i10} & -p_{i10} \cdot p_{i01} & -p_{i10} \cdot p_{i00} \\ \sigma_{13(i)} & \sigma_{23(i)} & p_{i01} \cdot q_{i01} & -p_{i01} \cdot p_{i00} \\ \sigma_{14(i)} & \sigma_{24(i)} & \sigma_{34(i)} & p_{i00} \cdot q_{i00} \end{bmatrix} \tag{2}$$

To measure the effect of the exposure *E* and the extraneous *C*, calculate respectively from Table 1 the following estimates for the exposure odds ratios of *E* unstratified and stratified by *C*:

$$\hat{R}_E = (n_{11+} \cdot n_{00+}) / (n_{10+} \cdot n_{01+}) \tag{3}$$

$$\hat{R}_{E|C=1} = (n_{111} \cdot n_{001}) / (n_{101} \cdot n_{011}) \tag{4}$$

and

$$\hat{R}_{E|C=0} = (n_{110} \cdot n_{000}) / (n_{100} \cdot n_{010}) \tag{5}$$

In addition, the Mantel-Haenszel summary odds ratio is given by (Mantel and Haenszel 1959)

$$\hat{R}_{E|MH} = (n_{111} \cdot n_{001} \cdot n_{++1}^{-1} + n_{110} \cdot n_{000} \cdot n_{++0}^{-1}) \cdot (n_{101} \cdot n_{011} \cdot n_{++1}^{-1} + n_{100} \cdot n_{010} \cdot n_{++0}^{-1})^{-1} \tag{6}$$

Now let the ratios of odds ratios (Eqs. 3-6) be defined respectively by

$$\hat{\phi}_{E|C} = \hat{R}_{E|MH} / \hat{R}_E \tag{7}$$

and

$$\hat{\phi}_{hmg} = \hat{R}_{E|C=1} / \hat{R}_{E|C=0} \tag{8}$$

where "hmg" in the subscript of Eq. 8 denote the word "homogeneous". If the estimated value of $\hat{\phi}_{E|C}$ is greater or less than 10% of the null value of unity, the extraneous variable *C* is said to be a confounder. However, this condition is only sufficient, but not necessary as other conditions will be given in the section of Discussion. Two strata are said to be heterogeneous if the estimated value of $\hat{\phi}_{hmg}$ is significantly different from the null value of unity; otherwise, it is said to be homogeneous.

Let $R_{1^{st}\,OI(i)}$ and $R_{2^{nd}\,OI}$ denote respectively the 1st and 2nd order (multiplicative) interaction between *E* and *C* and among *D*, *E* and *C* (Lee 2012). Then the estimates of these ratios are given respectively as follows:

$$\hat{R}_{1^{st}\,OI(i)} = (n_{i11} \cdot n_{i00}) / (n_{i10} \cdot n_{i01}) \tag{9}$$

and

$$\hat{R}_{2^{nd}\,OI} = (n_{111} \cdot n_{010} \cdot n_{100} \cdot n_{010}) / (n_{101} \cdot n_{011} \cdot n_{110} \cdot n_{000}) \tag{10}$$

Two variables *E* and *C* are said to be have a first order multiplicative interaction if the estimated value of Eq. 9 is significantly different from the null value of unity; otherwise, there does not exist multiplicative interaction between *E* and *C*. Three variables *D*, *E* and *C* are said to be have a second-order multiplicative interaction if the estimated value of Eq. 10 is significantly different from the null value of unity; otherwise, there does not exist second-order multiplicative interaction among *D*, *E* and *C*. The extraneous variable *C* is said to be an effect modifier if either the estimated value of Eqs. 8 or 10 are significantly different from the null value of unity; otherwise, *C* is not an effect modifier. By the way, it is easy to show that Eq. 8 equals Eq. 10.

In addition, let R_C denote a measure of the strength of confounding by the extraneous variable *C* (Miettinen 1972),

$$\hat{R}_C = (n_{110} \cdot n_{101} \cdot n_{100}^{-1} + n_{010} \cdot n_{001} \cdot n_{000}^{-1}) \cdot n_{00+} / (n_{10+} \cdot n_{01+}) \tag{11}$$

E and *C* are jointly misclassified

Suppose that *D* is not misclassified at all, but only *E* and *C* are jointly misclassified. There are two scenarios between *E* and *C* which have to be considered separately.

Scenario I: The classification errors of *E* and *C* are independent.

For this scenario, each cell misclassification probability is obtained as the product of the corresponding two row/column marginal misclassification probabilities. The false positive and false negative probabilities for $Y = E$ or C are defined as follows:

$$\gamma_{Y(i)} = \Pr(Y^* = 0 \mid Y = 1; D = i) \text{ and } \delta_{Y(i)} = \Pr(Y^* = 1 \mid Y = 0; D = i) \tag{12}$$

For $i = 1, 0$, let

$$p_{(i)} = [p_{i11}, p_{i10}, p_{i01}, p_{i00}]^T \tag{13a}$$

$$\hat{p}_{(i)} = [\hat{p}_{i11}, \hat{p}_{i10}, \hat{p}_{i01}, \hat{p}_{i00}]^T \tag{13b}$$

Thus, by using Eq. 12, for $i = 1, 0$

$$E(\hat{p}_{(i)}) = W_{I(i)} p_{(i)} \tag{14}$$

where the misclassification matrix $W_{I(i)}$ is given by (Barron 1977; Tzonou et al. 1986)

$$\begin{aligned} W_{I(i)} &= [w_{I(ijk)}] \\ &\equiv \begin{bmatrix} (1-\gamma_{E(i)})(1-\gamma_{C(i)}) & (1-\gamma_{E(i)})\delta_{C(i)} & (1-\gamma_{C(i)})\delta_{E(i)} & \delta_{E(i)}\delta_{C(i)} \\ (1-\gamma_{E(i)})\gamma_{C(i)} & (1-\gamma_{E(i)})(1-\delta_{C(i)}) & \delta_{E(i)}\gamma_{C(i)} & \delta_{E(i)}(1-\delta_{C(i)}) \\ \gamma_{E(i)}(1-\gamma_{C(i)}) & \gamma_{E(i)}\delta_{C(i)} & (1-\delta_{E(i)})(1-\gamma_{C(i)}) & (1-\delta_{E(i)})\delta_{C(i)} \\ \gamma_{E(i)}\gamma_{C(i)} & \gamma_{E(i)}(1-\delta_{C(i)}) & (1-\delta_{E(i)})\gamma_{C(i)} & (1-\delta_{E(i)})(1-\delta_{C(i)}) \end{bmatrix} \end{aligned} \tag{15}$$

By conditioning on that $\gamma_{Y(i)}$ and $\delta_{Y(i)}$ for $Y = E$ or C are known, the vector of bias-adjusted cell probability estimator (BACP) $\breve{p}_{I(i)} (= [\breve{p}_{i11(I)}, \breve{p}_{i10(I)}, \breve{p}_{i01(I)}, \breve{p}_{i00(I)}]^T)$ is then defined by

$$\breve{p}_{I(i)} = W_{I(i)}^{-1}\hat{p}_{(i)} = V_{I(i)}\hat{p}_{(i)} \tag{16}$$

where the inverse $V_{I(i)}$ of $W_{I(i)}$ is given by

$$\begin{aligned} V_{I(i)} &\equiv W_{I(i)}^{-1} \\ &= \Delta_{I(i)}^{-\frac{1}{2}} \cdot \begin{bmatrix} (1-\delta_{E(i)})(1-\delta_{C(i)}) & -\gamma_{C(i)}(1-\delta_{E(i)}) & -\gamma_{E(i)}(1-\delta_{C(i)}) & \gamma_{E(i)}\gamma_{C(i)} \\ -\delta_{C(i)}(1-\delta_{E(i)}) & (1-\gamma_{C(i)})(1-\delta_{E(i)}) & \gamma_{E(i)}\delta_{C(i)} & -\gamma_{E(i)}(1-\gamma_{C(i)}) \\ -\delta_{E(i)}(1-\delta_{C(i)}) & \delta_{E(i)}\gamma_{C(i)} & (1-\gamma_{E(i)})(1-\delta_{C(i)}) & -\gamma_{C(i)}(1-\gamma_{E(i)}) \\ \delta_{E(i)}\delta_{C(i)} & -\delta_{E(i)}(1-\gamma_{C(i)}) & -\delta_{C(i)}(1-\gamma_{E(i)}) & (1-\gamma_{E(i)})(1-\gamma_{C(i)}) \end{bmatrix} \end{aligned} \tag{17}$$

where $\Delta_{I(i)} \equiv \det(W_{I(i)}) = [(1-\gamma_{E(i)}-\delta_{E(i)})(1-\gamma_{C(i)}-\delta_{C(i)})]^2$.

In order for W to be invertible, the following constraints on its false positive and false negative rates for both exposure and confounding variables are imposed:

$$\gamma_{E(i)} + \delta_{E(i)} < 1 \text{ and } \gamma_{C(i)} + \delta_{C(i)} < 1 \tag{18}$$

Scenario II: The classification errors of E and C are not independent.

For this scenario, there are 16 possibly cross-classified conditional probabilities of E^* and C^* as follows: for fixed j', k', i, k = 1, 0

$$\lambda_{j'k'(i)}^{jk} = \Pr(E^* = j, C^* = k \mid E = j', C = k'; D = i) \tag{19}$$

where $\{\lambda_{j'k'(i)}^{jk}\}$, for j', k' = 1, 0, are required to satisfy the following identities:

$$\sum_{j,k=0}^{1} \lambda_{j'k'(i)}^{jk} = 1,\ 0 \le \lambda_{j'k'(i)}^{jk} \le 1 \tag{20}$$

Among the $\{\lambda_{j'k'(i)}^{jk}\}$, four are correctly classified and 12 are misclassification probabilities. Because the misclassification can go equally from one cell to another three cells, it is appropriate to assume that they all equal to one another, that is, $\theta_{1(i)} \equiv \lambda_{11(i)}^{10} = \lambda_{11(i)}^{01} = \lambda_{11(i)}^{00}$, $\theta_{2(i)} \equiv \lambda_{10(i)}^{11} = \lambda_{10(i)}^{01} = \lambda_{10(i)}^{00}$,

$\theta_{3(i)} \equiv \lambda_{01(i)}^{11} = \lambda_{01(i)}^{10} = \lambda_{01(i)}^{00}$, and $\theta_{4(i)} \equiv \lambda_{00(i)}^{11} = \lambda_{00(i)}^{10} = \lambda_{00(i)}^{01}$. Thus, the misclassification matrix is given by

$$W_{II(i)} = \begin{bmatrix} 1-\theta_{2(i)}-\theta_{3(i)}-\theta_{4(i)} & \theta_{2(i)} & \theta_{3(i)} & \theta_{4(i)} \\ \theta_{1(i)} & 1-\theta_{1(i)}-\theta_{3(i)}-\theta_{4(i)} & \theta_{3(i)} & \theta_{4(i)} \\ \theta_{1(i)} & \theta_{2(i)} & 1-\theta_{1(i)}-\theta_{2(i)}-\theta_{4(i)} & \theta_{4(i)} \\ \theta_{1(i)} & \theta_{2(i)} & \theta_{3(i)} & 1-\theta_{1(i)}-\theta_{2(i)}-\theta_{3(i)} \end{bmatrix} \tag{21}$$

In addition, the inverse matrix $V_{\text{II(i)}}$ of $W_{\text{II(i)}}$ is given by

$$\begin{aligned} V_{II(i)} &\equiv W_{II(i)}^{-1} \\ &= \Delta_{II(i)}^{-\frac{1}{3}} \begin{bmatrix} 1-\theta_{1(i)}-2\theta_{2(i)} & -\theta_{1(i)} & -\theta_{1(i)} & -\theta_{1(i)} \\ -\theta_{2(i)} & 1-\theta_{2(i)}-2\theta_{4(i)} & -\theta_{2(i)} & -\theta_{2(i)} \\ -\theta_{3(i)} & -\theta_{3(i)} & 1-2\theta_{1(i)}-\theta_{3(i)} & -\theta_{3(i)} \\ -\theta_{4(i)} & -\theta_{4(i)} & -\theta_{4(i)} & 1-2\theta_{3(i)}-\theta_{4(i)} \end{bmatrix} \end{aligned} \tag{22}$$

where $\Delta_{II(i)} = \det(W_{II(i)}) = (1-\theta_{1(i)}-\theta_{2(i)}-\theta_{3(i)}-\theta_{4(i)})^3$. For this scenario, the BACP estimator is given by

$$\breve{p}_{II(i)} = W_{II(i)}^{-1}\hat{p}_{(i)} = V_{II(i)}\hat{p}_{(i)} \tag{23}$$

The misclassification probabilities ((γ_Y, δ_Y) or $(\theta_{1(i)}, \theta_{2(i)}, \theta_{3(i)}, \theta_{4(i)})$) are said to be feasible if the misclassification matrix ($W_{\text{I(i)}}$ or $W_{\text{II(i)}}$) is nonsingular, or equivalently, its determinant is nonzero. The BACP estimator (Eqs. 16 or 23) is said to be admissible if every component of its vector is nonnegative and their sum equals to the total probability one. In theory, it is possible to find the admissibility constraints which are required to be imposed on the misclassification probability. Yet, because it does not yield inequalities as neat as that of case-control studies with a single exposure variable (Lee 2009), it is therefore omitted here. Nevertheless, the admissibility constraints can be checked in practical applications by taking a case-by-case approach as illustrated by the example in the next section.

From Eq. 16, for j, k = 1, 0

$$\breve{n}_{ijk} \equiv n_{(i)} \cdot \breve{p}_{ijk} \tag{24}$$

By substituting Eqs. 16 or 23 into Eqs. 3-11, obtain the corresponding bias-adjusted estimates $\breve{R}_E, \breve{R}_{E|C=1}, \breve{R}_{E|C=0}, \breve{R}_{E|MH}$, $\breve{\phi}_E$, $\breve{\phi}_I$, $\breve{R}_{1^{st}\,OI\,(i)}$, $\breve{R}_{2^{nd}\,OI}$, and $\breve{R}_C$. Note that the crude estimators are merely a special case of the lower end of the bias-adjusted one when its false positive and false negative rates are all zero.

Example

The data in Table 2 are taken from Crane et al. (1989). A case-control study was conducted to examine the hypothesis that fenoterol by metered dose inhaler increases the risk of death in patients with asthma. Cases were drawn from the National Asthma Mortality Survey which identified all asthma deaths in New Zealand from August 1981 to July 1983. Of the 271 asthma deaths identified in the survey, 125 occurred in patients aged 5-45 years, and these formed the case group. For each case, 4 controls, matched for age and ethnic group, were selected from asthma admissions to hospitals to which the cases themselves would have been admitted, had they survived. Controls were obtained for 124 out of the 125 cases. 7 cases were subsequently excluded because they died after admission to hospitals. Therefore the analysis pertains to 117 cases and 468 matched controls.

In terms of symbols, the disease, exposure and extraneous variables are given as follows:

D = asthma death (= 1 if outpatient deaths, = 0 if hospitalized controls),
E = use of prescribed fenoterol (= 1 if yes, = 0 if no),
C = use of corticosteroids (= 1 if used, = 0 if not used).

If the data are not misclassified, the crude estimators with its 95% confidence interval (*CI*) are obtained by using Eqs. 3-11 and A.1-A.15 in the appendix: $\hat{R}_E = 1.55$ (95% *CI*: 1.03 – 2.33), $\hat{R}_{E|C=1} = 6.45$ (95% *CI*: 2.56 – 16.3), $\hat{R}_{E|C=0} = 0.96$ (95% *CI*: 0.59 – 1.55), $\hat{R}_{E|MH} = 1.53$ (95% *CI*: 1.24 – 1.87), $\hat{\phi}_{E|C} = 0.98$ (95% *CI*: 0.62 – 1.55), $\hat{\phi}_{hmg} = 6.73$ (95% *CI*: 2.37 – 19.1), $\hat{R}_{1^{st}\,OI\,(1)} = 5.46$ (95% *CI*: 2.13 – 14.0), $\hat{R}_{1^{st}\,OI\,(0)} = 0.81$ (95% *CI*: 0.52 – 1.27), $\hat{R}_{2^{nd}\,OI} = \hat{\phi}_{hmg}$ (95% *CI*: same as $\hat{\phi}_{hmg}$), and $\hat{R}_C = 1.34$ (95% *CI*:1.03 – 2.33).

As was pointed out by O'Donnell et al (1989) who reviewed the data for the group allegedly at highest risk, some of 20 such fatal cases which were recorded by Crane et al might not use prescribed fenoterol as reported by the general practitioner. In 7 of the 18 cases, the beta-agonist in use immediately before death was not or might not have been fenoterol. Hence, the data used by Crane et al were likely to be misclassified.

Suppose that the data are misclassified. The proposed bias-adjusted estimator can be used, namely, replacing n_{ijk} in Eqs. 3-11 by the values of Eq. 24 to account for the misclassification bias. However, before calculating $\breve{n}_{ijk}$ in Eq. 24, calculate the misclassification probabilities. Here, the idea of counterfactual thinking was employed in creating the correctly classified table to serve as a gold standard for calculating the misclassification probability (Epstude & Roese 2008). Recall that the actually observed table is the only concrete source of information. Therefore, the observed table is taken as a factual one. Then, the correctly classified table is nothing but a counterfactual (*CF*) table corresponding to the factual (observed misclassified) table. *CF* tables are said to be feasible (or admissible) if the misclassification matrix associated with the calculated misclassification probabilities is nonsingular, namely, its determinant is nonzero (or if the bias-adjusted cell probability estimators (Eqs. 16 or 23) are admissible).

At first, the construction of 20 *CF* tables was tried. However, only 8 counterfactual models for cases and controls were listed here. Even among these 8 models, only 2 models (models 4 and 5, boldface in Table 3a) for cases under scenario I were admissible. For all other models either the 3rd component of the BACP had a negative value (*CF* tables 1 and 8) or the sum of the all components of the BACP estimator did not equal one (*CF* tables 2-3 and 6-7) (Table 3a, column 5). Even worse, none under scenario II for cases were admissible, because either the 3rd component of the BACP estimator had a negative value (*CF* tables 1-2 and 7-8) or the sum of all four components of BACP estimator did not equal one (*CF* tables 3-6) (Table 3a, column 8). For controls, only *CF* tables 3-6 were admissible under either scenario I or II (boldface in Table 3b).

Table 2: A case-control study for the asthma deaths in New Zealand

$D = 1$		$C^* = 1$	$C^* = 0$	Total
(Cases)	$E^* = 1$	26	34	60
	$E^* = 0$	7	50	57
	Total	33	84	117
	$E^* = 1$	38	151	189
$D = 0$ (Controls)	$E^* = 0$	66	213	279
	Total	**104**	**364**	**468**

Table 3: Counterfactual tables with its false positive and false negative rates, determinant of the misclassification matrix under two scenarios for

(a) Cases

		Scenario I			**Scenario II**		
CF table	(n_{111}, n_{110}, n_{101}, n_{100})	($\gamma_{E(1)}$, $\delta_{E(1)}$, $\gamma_{C(1)}$, $\delta_{C(1)}$)	$\Delta_{I(1)}$	($\bar{p}_{I(111)}$, $\bar{p}_{I(110)}$, $\bar{p}_{I(101)}$, $\bar{p}_{I(100)}$)	($\theta_{1(1)}$, $\theta_{2(1)}$, $\theta_{3(1)}$, $\theta_{4(1)}$)	$\Delta_{II(1)}$	($\bar{p}_{II(111)}$, $\bar{p}_{II(110)}$, $\bar{p}_{II(101)}$, $\bar{p}_{II(100)}$)
1	(30, 38, 11, 38)	(0.06, 0.08, 0.11, 0.05)	0.53	(0.23, 0.28, -0.005, 0.46)	(0.02, 0.02, 0.07, 0.05)	0.59	(0.23, 0.29, -0.02, 0.38)
2	(29, 37, 10, 41)	(0.05, 0.06, 0.08, 0.04)	0.62	(0.23, 0.29, 0.01, 0.45)	(0.02, 0.01, 0.06, 0.03)	0.67	(0.23, 0.29, -0.001, 0.39)
3	(28, 36, 9, 44)	(0.03, 0.04, 0.06, 0.02)	0.73	(0.22, 0.29, 0.03, 0.44)	(0.01, 0.01, 0.04, 0.02)	0.77	(0.22, 0.29, 0.02, 0.40)
4	(27, 35, 8, 47)	(0.02, 0.02, 0.03, 0.01)	0.86	**(0.22, 0.29, 0.05, 0.43)**	(0.006, 0.005, 0.02, 0.01)	0.87	(0.224, 0.293, 0.039, 0.416)
5	(25, 33, 6, 53)	(0.02, 0.02, 0.03, 0.01)	0.85	**(0.22, 0.29, 0.04, 0.43)**	(0.007, 0.005, 0.03, 0.01)	0.87	(0.224, 0.294, 0.035, 0.415)
6	(24, 32, 5, 56)	(0.03, 0.03, 0.06, 0.02)	0.72	(0.22, 0.29, 0.03, 0.44)	(0.01,0.01, 0.06, 0.02)	0.73	(0.23, 0.30, 0.003, 0.40)
7	(23, 31, 4, 59)	(0.05, 0.05, 0.10, 0.03)	0.60	(0.23, 0.28, 0.006, 0.45)	(0.02,0.02, 0.09, 0.03)	0.61	(0.23, 0.31, -0.04, 0.38)
8	(22, 30, 3, 62)	(0.07, 0.07, 0.14, 0.05)	0.50	(0.23, 0.28, -0.02, 0.46)	(0.03,0.02, 0.13, 0.04)	0.48	(0.24, 0.32, -0.10, 0.35)

Table 3 Continued

(b) Controls

CF table	(n_{011}, n_{010}, n_{001}, n_{000})	**Scenario I** ($\gamma_{E(0)}$, $\delta_{E(0)}$, $\gamma_{C(0)}$, $\delta_{C(0)}$)	$\Delta_{I(0)}$	($\bar{p}_{I(011)}$, $\bar{p}_{I(010)}$, $\bar{p}_{I(001)}$, $\bar{p}_{I(000)}$)	**Scenario II** ($\theta_{1(0)}$, $\theta_{2(0)}$, $\theta_{3(0)}$, $\theta_{4(0)}$)	$\Delta_{II(0)}$	($\bar{p}_{II(011)}$, $\bar{p}_{II(010)}$, $\bar{p}_{II(001)}$, $\bar{p}_{II(000)}$)
1	(42, 155 70, 201)	(0.02, 0.01, 0.04, 0.01)	0.84	(0.07, 0.32, 0.13, 0.46)	(0.02, 0.004, 0.01, 0.01)	0.88	(0.07, 0.33, 0.13, 0.45)
2	(41, 154, 69, 204)	(0.02, 0.01, 0.03, 0.01)	0.88	(0.07, 0.32, 0.13, 0.46)	(0.01, 0.003, 0.01, 0.01)	0.91	(0.07, 0.32, 0.13, 0.46)
3	(40, 153, 68, 207)	(0.01, 0.01, 0.02, 0.01)	0.92	**(0.08, 0.32, 0.14, 0.46)**	(0.01, 0.002, 0.005, 0.005)	0.94	**(0.07, 0.32, 0.14, 0.46)**
4	(39, 152, 67, 210)	(0.005,0.004, 0.01, 0.003)	0.96	**(0.08, 0.32, 0.14, 0.46)**	(0.004, 0.001, 0.003, 0.002)	0.97	**(0.08, 0.32, 0.14, 0.46)**
5	(37, 150, 65, 216)	(0.005,0.004, 0.01, 0.003)	0.96	**(0.08, 0.32, 0.14, 0.46)**	(0.004, 0.001, 0.003, 0.002)	0.97	**(0.08, 0.32, 0.14, 0.46)**
6	(36, 149, 64, 219)	(0.01, 0.007, 0.02, 0.005)	0.92	**(0.08, 0.32, 0.14, 0.46)**	(0.01, 0.002, 0.005, 0.005)	0.94	**(0.07, 0.32, 0.14, 0.46)**
7	(35, 148, 63, 222)	(0.02, 0.01, 0.03, 0.01)	0.88	(0.07, 0.32, 0.13, 0.46)	(0.01, 0.003, 0.01, 0.001)	0.91	(0.07, 0.33, 0.13, 0.46)
8	(34, 147, 62, 225)	(0.02, 0.01, 0.04, 0.01)	0.84	(0.07, 0.32, 0.13, 0.46)	(0.02, 0.004, 0.01, 0.01)	0.88	(0.06, 0.33, 0.13, 0.46)

Table 4: Estimated values of bias-adjusted estimators for all statistics (Eqs. 3-11) with its 95% CI for selected admissible counterfactual tables

Test (95% CI)	***CF* table: ($case_l$, $control_l$)**			
	(#4, #3)	(#4, #4)	(#4, #5)	(#4, #6)
$\breve{R}_E$	1.59 (1.22 – 2.06)	1.58 (1.22 – 2.05)	1.58 (1.22 – 2.06)	1.59 (1.22 – 2.06)
$\breve{R}_{E\|MH}$	1.554 (1.22 – 1.98)	1.551 (1.22 – 1.97)	1.551 (1.21 – 1.98)	1.554 (1.22 – 1.97)
$\breve{R}_{E\|C=1}$	8.72 (3.00 – 25.3)	8.62 (2.98 – 25.0)	8.63 (2.98 – 25.0)	8.72 (2.99 – 25.4)
$\breve{R}_{E\|C=0}$	0.94 (0.35 – 2.58)	0.94 (0.35 – 2.57)	0.94 (0.35 – 2.57)	0.94 (0.35 – 2.58)

Table 4 Continued

Test (95% CI)	*CF* table: ($case_l$, $control_l$)			
$\breve{\phi}_{E\|C}$	0.98 (0.23 – 4.23)	0.98 (0.23 – 4.23)	0.98 (0.23 – 4.23)	0.98 (0.23 – 4.25)
$\breve{\phi}_{hmg}$	9.23 (2.22 – 38.3)	9.16 (2.21 – 38.0)	9.29 (2.21 – 38.1)	9.36 (2.21 – 38.5)
$\breve{R}_{1^{st}\,OI\,(1)}$	7.39 (1.97 – 27.6)	7.39 (1.97 – 27.6)	7.39 (1.97 – 27.6)	7.39 (1.97 – 27.6)
$\breve{R}_{1^{st}\,OI\,(0)}$	0.8 (0.47 – 1.36)	0.81 (0.48 – 1.37)	0.81 (0.47 – 1.38)	0.8 (0.46 – 1.38)
$\breve{R}_{2^{nd}\,OI}$	9.23 (2.22 – 38.3)	9.16 (2.21 – 38.0)	9.16 (2.21 – 38.1)	9.24 (2.21 – 38.5)
$\breve{R}_{C}$	1.58 (1.04 – 2.39)	1.6 (1.10 – 2.32)	1.6 (1.09 – 2.33)	1.58 (1.07 – 2.31)

After getting all possible combinations from admissible *CF* tables under scenario I for cases and controls, the bias-adjusted values of Eqs. 3-11 were computed for all 8 combinations. Only 4 combinations were listed here because the results from the other 4 combinations were similar; hence it was omitted to save space (table 4). On the one hand, the bias-adjusted values for the unstratified exposure odds ratio, the Mantel-Haenezel summary odds ratios, the odds ratio for the stratum without the presence of *C*, the 1st order interaction for controls are almost unchanged. On the other hand, the bias-adjusted values for the stratified odds ratio with the presence of *C* and the 1st order interaction for cases were 35% higher than the crude estimator. Similarly, the bias-adjusted value for the 2nd order multiplicative interaction was 36% higher than the crude estimator.

Discussion

This is a study on the effect of joint misclassification of exposure and extraneous (confounding) variables on the association among the disease, exposure and confounding variables. Through the use of counterfactual tables as a gold standard, a sensitivity analysis was conducted to examine effects on various measures used in analyzing 2 × 2 × 2 tables. Both Cox & Elwood (1991) and Walker & Lanes (1991) also used the same data set to investigate the issue of misclassification. But, they only considered the effect of the confounder misclassification.

Some comments are appropriate to be given below:

1. Two scenarios concerning the joint misclassification were considered, that is, the classification errors of exposure and confounding variables are independent (scenario I) or not (scenario II). It turned out that results were very different for cases and controls. Under scenario I, there were 2 and 4 admissible *CF* tables for cases and controls respectively. It was noticed that no admissible *CF* tables exist, once the false positive or negative rates were greater than 0.02. Under scenario II none and 4 *CF* tables were available for cases and controls. Similarly, no admissible *CF* tables exist for controls, once the false positive or negative rates were greater than 0.01. Evidently, the existence of admissible *CF* tables depends on the structure of their collected data for cases and controls.

2. From the result of this study, the effect of joint misclassification of the exposure and confounding variables varies. It depends on which statistics is used to measure the effect (table 4). For example, although the value of $\breve{R}_E$ is just a little larger than that of $\hat{R}_E$, it implies that the bias-adjusted estimator is significantly greater than one because its lower bound of 95% *CI* moves further from one than that of $\hat{R}_E$. Its 95% *CI* becomes widened than that of $\hat{R}_E$, even though the values of $\breve{R}_{E|MH}$ are approximately the same as that of $\hat{R}_{E|MH}$. The values of $\breve{R}_{E|C=1}$, $\breve{R}_{1^{st}OI(1)}$ and $\breve{\phi}_{hmg}$ are much larger than that of the corresponding crude estimators $\hat{R}_{E|C=1}$, $\hat{R}_{1^{st}OI(1)}$ and $\hat{\phi}_{hmg}$. The value of $\breve{R}_C$ which is greater than that of $\hat{R}_C$ indicates that the strength of confounding by the use of oral corticosteroids is at least 1.3 times of the group without using it. Lastly, the effect of joint misclassification of *E* and *C* on the measure of $R_{E|C=0}$, $R_{1^{st}OI(0)}$, and $\phi_{E|C}$ is almost negligible.

3. Advantages in using counterfactual tables to conduct the sensitivity analysis are many folds. First, it solves the problem of finding a gold standard in order to calculate the misclassification probability. Second, the assumption of nondifferential misclassification is not needed on any factor under study. For example, if the exposure factor is not misclassified, all that has to be done is to keep the marginal totals for

the exposure factor fixed in selecting counterfactual tables. Third, all results are strictly obtained from the collected data. Hence, the drawn conclusion is real data-based rather than the hypothetical data used by other authors (Greenland 1980; Greenland & Robins 1985).

4. Although the extraneous variable *C* (the use of corticosteroid) was not judged as a confounder by the estimated value of $\hat{\phi}_{E|C} = 0.98$, it was shown to be a confounder by indication (Psaty et al. 1999) or by association with the death and the exposure (Miettinen 1974) or by not being equally distributed (lack of comparability) in the categories of the exposure variable (the use of prescribed fenoterol) (Miettinen 1985). Further, it was shown to be an effect modifier by the statistics of $\hat{\phi}_{hmg}$ or $\hat{R}_{2^{nd}OI}$. Once an effect modification is present, whether *C* is a confounder becomes not an issue. Rather, stratum-specific odds ratio estimates ($\hat{R}_{E|C=i}$) should be reported because summary estimates do not convey information on the pattern of variation of stratum-specific estimates. For other references on the confounder, please see Wickramaratne & Holford (1987), Weinberg (1993), and Yiostalo & Knuuttila (2006).

Acknowledgements

All numerical calculations were done with the use of the EXCEL spreadsheet.

References

Agresti, A. (2002). *Categorical Fata Analysis*. (2nd ed.) New York: Wiley.

Ahlbom, A. & Steinbeck, G. (1992). Aspects of misclassification of confounding factors. *American Journal Industrial Medicine, 21*, 107-112.

Axelson, O. (1978). Aspects on confounding in occupational health epidemiology. (Letter). *Scandinavian Journal of Work, Environment, & Health, 4*, 98-102.

Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics, 33*, 414-418.

Cox, B., & Elwood, J. M. (1991). The effect on the stratum-specific odds ratios of nondifferential misclassification of a dichotomous covariate. *American Journal of Epidemiology, 133*, 202-207.

Crane, J., Pearce, N., Flatt, A., Burgess, C., Jackson, R., Kwong, T., Ball, M., & Beasley, R. (1989). Prescribed fenterol and death from asthma in New Zealand, 1981-1983: Case-Control study. *Lancet, 1*, 917-922.

Epstude, K., & Roese, N. (2008). The functional theory of counterfactual thinking. *Personality And Social Psychology Review*, *12*(2), 168-192. doi:10.1177/1088868308316091

Fleiss, J., Levin, B., & Paik, M.C. (2003). *Statistical Methods for Rates and Proportions,* 3rd edition. John Wiley and Sons, Inc., New York.

Fung, K. Y., & Howe, G. R. (1984). Methodological issues in case-control studies III: - The effect of joint misclassification of risk factors and confounding factors upon estimation and power. *International Journal of Epidemiology, 13,* 366-370.

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology, 112*, 564-569.

Greenland, S., & Robins, J. M. (1985). Confounding and misclassification. *American Journal of Epidemiology*, *122,* 495-506.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall, Boca Raton, FL.

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. John Wiley and Sons, Inc., New York.

Kupper, L. L. (1984). Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *American Journal of Epidemiology*, *120*, 643-648.

Lee, T-S. (2009). Bias-adjusted exposure odds ratio for misclassified data. *The Internet Journal of Epidemiology*, *6*(*2*). Accessed from http://www.ispub.com/journal/the-internet-journal-of-epidemiology/volume-6-number-2/bias-adjusted-exposure-odds-ratio-for-misclassification-data-1.html

Lee, T-S. (2013). The effect on the interaction by the joint misclassification of two exposure factors in contingency tables. A special issue on Environmental Health Statistics: Epidemiology, Toxicology and Related Issues, *Journal of the Indian Society of Agricultural Statistics, 67,* 1-9.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Miettinen, O. S. (1972). Components of the crude risk ratio. *American Journal of Epidemiology, 96,* 168-172.

Miettinen, O. S. (1974). Confounding and effect-modification. *American Journal of Epidemiology., 100,* 350-353.

Miettinen, O. S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine.* New York: Wiley.

O'Donnell, T. V., Rea, H. H., Holst, & Sears, M. R. (1989). Fenoterol and fatal asthma. *The Lancet, 1,* 1070-1071.

Psaty, B. M., Koepsell, T. D., Lin, D., Weiss, N. S., Siscovick, D. S., Rosendaal, F. R., Pahor, M., & Furberg, C. D. (1999). Assessment and control for confounding by indication in observational studies. *Journal of the American Geriatrics Society, 47,* 749-754.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology*, 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Savitz, D. A., & Baron, A. E. (1989). Estimating and correcting for confounder misclassification. *American Journal of Epidemiology, 129,* 1062-1071.

Tzonou, A., Kaldor, J., Smith, P.G., Day, N.E., & Trichopoulos, D. (1986). Misclassification in case-control studies with two dichotomous risk factors. *Rev d'Epidém et de Santé Publ, 34,* 10-17.

Walker, A. M. (1985). Misclassified confounders. *American Journal of Epidemiology., 122,* 921-922

Walker, A. M., & Lanes, S. F. (1991). Misclassification of covariates. *Statistics in Medicine., 10,* 1181-1196.

Weinberg, C. R. (1993). Toward a clearer definition of confounding. *American Journal of Epidemiology, 137,* 1-8.

Wickramaratne, P. J. & Holford, T. R. (1987). Confounding in epidemiologic studies: The adequacy of the control group as a measure of confounding. *Biometrics, 43,* 751-765.

Yiostalo, P. V. & Knuuttila, M. L. (2006). Confounding and effect modification: possible explanation for variation in the results on the association between oral and systemic diseases. *Journal of Clinical Periodontology, 33,* 104-108.

Appendix A

By adapting the variance formula obtained in Lee (2013), the variance formula is readily given for the crude and bias-adjusted estimators for the 1st order (multiplicative) interaction as follows:

$$Var(\ln(\hat{R}_{1^{st}OI(i)})) = n_{(i)}^{-1} \cdot a_{(i)}^{T} \Sigma_{(i)} a_{(i)} \tag{A1}$$

where $a_{(i)} = (p_{i11}^{-1}, -p_{i10}^{-1}, -p_{i01}^{-1}, p_{i00}^{-1})^{T}$.

$$Var(\ln(\breve{R}_{1^{st}OI(i)}) = n_{(i)}^{-1} \cdot \dot{a}_{(i)}^{T} \Sigma_{(i)} \dot{a}_{(i)} \tag{A2}$$

where $\dot{a}_{(i)} = (v_{11(.;i)}\hat{p}_{i11}, -v_{22(.;i)}\hat{p}_{i10}, -v_{33(.;i)}\hat{p}_{i01}, v_{44(.;i)}\hat{p}_{i00})^{T}$, $\{v_{jj(.;i)}\}$ are the j[th] diagonal entry of the inverse matrix $V_{.(i)} = W_{.(i)}^{-1}$, $W_{.(i)} = W_{I(i)}$ or $W_{II(i)}$, $\hat{p}_{(i)} = V_{.(i)} p_{(i)}$, and $\Sigma_{(i)}$ is given by Eq. 2.

The variance formula is also readily given for the crude and bias-adjusted estimators for the 2nd order (multiplicative) interaction as follows:

$$Var(\ln(\hat{R}_{2^{nd}OI})) = \sum_{i=0}^{1} Var(\ln(\hat{R}_{1^{st}OI(i)})) \tag{A3}$$

$$Var(\ln(\breve{R}_{2^{nd}OI})) = \sum_{i=0}^{1} Var(\ln(\breve{R}_{1^{st}OI(i)})) \tag{A4}$$

Similarly, obtain the variance for the crude and bias-adjusted odds ratio ignoring the confounding factor C to be given respectively by using the delta method which is given by Eq. 14.4 in Agresti (2002):

$$Var(\ln(\hat{R}_{E})) = \sum_{i=0}^{1} n_{(i)}^{-1} \cdot b_{(i)}^{T} \Sigma_{(i)} b_{(i)} \tag{A5}$$

where $b_{(1)} = (p_{11+}^{-1}, p_{11+}^{-1}, -p_{10+}^{-1}, -p_{10+}^{-1})^{T}$, and $b_{(0)} = (-p_{01+}^{-1}, -p_{01+}^{-1}, p_{00+}^{-1}, p_{00+}^{-1})^{T}$.

The variance of $\ln(\breve{R}_{E})$ is given by

$$Var(\ln(\breve{R}_{E})) = \sum_{i=0}^{1} n_{(i)}^{-1} \dot{b}_{(i)}^{T} \Sigma_{(i)} \dot{b}_{(i)} \tag{A6}$$

where $\dot{b}_{(1)} = (v_{11(.;1)}\hat{p}_{11+}^{-1}, v_{22(.;1)}\hat{p}_{11+}^{-1}, -v_{33(.;1)}\hat{p}_{10+}^{-1}, -v_{44(.;1)}\hat{p}_{10+}^{-1})^T$, $\dot{b}_{(0)} = (-v_{11(.;0)}\hat{p}_{01+}^{-1}, -v_{22(.;0)}\hat{p}_{01+}^{-1}, v_{33(.;0)}\hat{p}_{00+}^{-1}, v_{44(.;0)}\hat{p}_{00+}^{-1})^T$, and $\hat{p}_{(i)}$ is defined in Eq. A2.

The variances of the crude and bias-adjusted odds ratio stratified by the confounder *C* are given respectively by

$$Var(\ln(\hat{R}_{E|C=1})) = \sum_{i=0}^{1} n_{(i)}^{-1} \cdot c_{i|C=1}\Sigma_{(i)}c_{(i)|C=1}^T \tag{A7}$$

(A8)

where $c_{(i)|C=1} = (p_{i11}^{-1}, 0, -p_{i01}^{-1}, 0)^T$, and $c_{(i)|C=0} = (0, -p_{i10}^{-1}, 0, p_{i00}^{-1})^T$.

(A9)

$$Var(\ln(\breve{R}_{E|C=1})) = \sum_{i=0}^{1} n_{(i)}^{-1}\dot{c}_{(i)|C=1}^T\Sigma_{(i)}\dot{c}_{(i)|C=1} \tag{A10}$$

where $\dot{c}_{(i)|C=1} = (v_{11(.;i)}\hat{p}_{i11}^{-1}, 0, -v_{33(.;i)}\hat{p}_{i01}^{-1}, 0)^T$, and $\dot{c}_{(i)|C=0} = (0, -v_{22(.;i)}\hat{p}_{i10}^{-1}, 0, v_{44(.;i)}\hat{p}_{i00}^{-1})^T$.

The variance of the crude Mantel-Haenszel summary odds ratio is given by

$$Var(\ln(\hat{R}_{E|MH})) = \sum_{i=0}^{1} n_{(i)}^{-1} \cdot d_{(i)}^T\Sigma_{(i)}d_{(i)} \tag{A11}$$

where $d_{(i)} = (d_{1(i)}, d_{2(i)}, d_{3(i)}, d_{4(i)})^T$ and each component in $d_{(i)}$ is given by

$$\begin{aligned}
d_{1(i=1)} &= (p_{++0}/p_{++1})[\rho_1 p_{001}(p_{101}+p_{011}+p_{001})+\rho_0 p_{101}p_{011}],\\
d_{2(i=1)} &= (p_{++1}/p_{++0})[\rho_1 p_{000}(p_{100}+p_{010}+p_{000})+\rho_0 p_{100}p_{010}],\\
d_{3(i=1)} &= -(p_{++0}/p_{++1})[\rho_1 p_{111}p_{001}+\rho_0 p_{011}(p_{111}+p_{011}+p_{001})],\\
d_{4(i=1)} &= -(p_{++1}/p_{++0})[\rho_1 p_{110}p_{000}+\rho_0 p_{010}(p_{110}+p_{010}+p_{000})],\\
d_{1(i=0)} &= -(p_{++0}/p_{++1})[\rho_1 p_{111}p_{001}+\rho_0 p_{101}(p_{111}+p_{101}+p_{001})],\\
d_{2(i=0)} &= -(p_{++1}/p_{++0})[\rho_1 p_{110}p_{000}+\rho_0 p_{100}(p_{110}+p_{100}+p_{000})],\\
d_{3(i=0)} &= (p_{++0}/p_{++1})[\rho_1 p_{111}(p_{111}+p_{101}+p_{001})+\rho_0 p_{101}p_{011}],\\
d_{4(i=0)} &= (p_{++1}/p_{++0})[\rho_1 p_{110}(p_{110}+p_{100}+p_{010})+\rho_0 p_{100}p_{010}],
\end{aligned}$$

$\rho_1 \equiv (p_{++0}p_{111}p_{001} + p_{++1}p_{110}p_{000})^{-1}$,
$\rho_0 \equiv (p_{++0}p_{101}p_{011} + p_{++1}p_{100}p_{010})^{-1}$.

The variance for the BACP of the Mantel-Haenszel summary odds ratio is given by

$$Var(\ln(\breve{R}_{E|MH})) = \sum_{i=0}^{1} n_{(i)}^{-1} \cdot \dot{d}_{(i)} \Sigma_{(i)} \dot{d}_{(i)}^T \tag{A12}$$

where $\dot{d}_{(i)} = [\dot{d}_{1(i)}, \dot{d}_{2(i)}, \dot{d}_{3(i)}, \dot{d}_{4(i)}]^T$ and each component in $\dot{d}_{(i)}$ is given by

$\dot{d}_{1(i=1)} = v_{11(.;1)}(\hat{p}_{++0} / \hat{p}_{++1})[\dot{\rho}_1 \hat{p}_{001}(\hat{p}_{101} + \hat{p}_{011} + \hat{p}_{001}) + \dot{\rho}_0 \hat{p}_{101}\hat{p}_{011}]$,
$\dot{d}_{2(i=1)} = v_{22(.;1)}(\hat{p}_{++1} / \hat{p}_{++0})[\dot{\rho}_1 \hat{p}_{000}(\hat{p}_{100} + \hat{p}_{010} + \hat{p}_{000}) + \dot{\rho}_0 \hat{p}_{100}\hat{p}_{010}]$,
$\dot{d}_{3(i=1)} = -v_{33(.;1)}(\hat{p}_{++0} / \hat{p}_{++1})[\dot{\rho}_1 \hat{p}_{111}\hat{p}_{001} + \dot{\rho}_0 \hat{p}_{011}(\hat{p}_{111} + \hat{p}_{011} + \hat{p}_{001})]$,
$\dot{d}_{4(i=1)} = -v_{44(.;1)}(\hat{p}_{++1} / \hat{p}_{++0})[\dot{\rho}_1 \hat{p}_{110}\hat{p}_{000} + \dot{\rho}_0 \hat{p}_{010}(\hat{p}_{110} + \hat{p}_{010} + \hat{p}_{000})]$,
$\dot{d}_{1(i=0)} = -v_{11(.;0)}(\hat{p}_{++0} / \hat{p}_{++1})[\dot{\rho}_1 \hat{p}_{111}\hat{p}_{001} + \dot{\rho}_0 \hat{p}_{101}(\hat{p}_{111} + \hat{p}_{101} + \hat{p}_{001})]$,
$\dot{d}_{2(i=0)} = -v_{22(.;0)}(\hat{p}_{++1} / \hat{p}_{++0})[\dot{\rho}_1 \hat{p}_{110}\hat{p}_{000} + \dot{\rho}_0 \hat{p}_{100}(\hat{p}_{110} + \hat{p}_{100} + \hat{p}_{000})]$,
$\dot{d}_{3(i=0)} = v_{33(.;0)}(\hat{p}_{++0} / \hat{p}_{++1})[\dot{\rho}_1 \hat{p}_{111}(\hat{p}_{111} + \hat{p}_{101} + \hat{p}_{001}) + \dot{\rho}_0 \hat{p}_{101}\hat{p}_{011}]$,
$\dot{d}_{4(i=0)} = v_{44(.;0)}(\hat{p}_{++1} / \hat{p}_{++0})[\dot{\rho}_1 \hat{p}_{110}(\hat{p}_{110} + \hat{p}_{100} + \hat{p}_{010}) + \dot{\rho}_0 \hat{p}_{100}\hat{p}_{010}]$,
$\dot{\rho}_1 \equiv (\hat{p}_{++0}\hat{p}_{111}\hat{p}_{001} + \hat{p}_{++1}\hat{p}_{110}\hat{p}_{000})^{-1}$,
$\dot{\rho}_0 \equiv (\hat{p}_{++0}\hat{p}_{101}\hat{p}_{011} + \hat{p}_{++1}\hat{p}_{100}\hat{p}_{010})^{-1}$.

The variance of $\ln(\hat{R}_C)$ is given by

$$Var(\ln(\hat{R}_C)) = \sum_{i=0}^{1} n_{(i)}^{-1} \cdot e_{(i)}^T \Sigma_{(i)} e_{(i)} \tag{A13}$$

where each component of the vector $e_{(i)} = (e_{1(i)}, e_{2(i)}, e_{3(i)}, e_{4(i)})^T$ is given by

$e_{1(1)} = 0$, $e_{2(1)} = \rho p_{101}p_{000}$, $e_{3(1)} = \rho p_{10+}^{-1} p_{100}(p_{110}p_{000} - p_{010}p_{001})$,
$e_{4(1)} = \rho p_{10+}^{-1} p_{100}^{-1}(p_{110}p_{000}p_{101}^2 - p_{010}p_{001}p_{100}^2)$; $e_{1(0)} = -p_{01+}^{-1}$,
$e_{2(0)} = \rho p_{01+}^{-1}(p_{100}p_{001}p_{011} - p_{000}p_{110}p_{101})$,
$e_{3(0)} = \rho[p_{000}(p_{110}p_{101} + p_{100}p_{010}) + 2p_{100}p_{010}p_{001}]$,

$e_{4(0)} = \rho p_{00+}^{-1} p_{000} p_{001} (p_{110} p_{000}^2 - p_{100} p_{010} p_{001})$; $\rho = (p_{000} p_{110} p_{101} + p_{100} p_{010} p_{001})^{-1}$.

The variance of $\ln(\breve{R}_C)$ is given by

$$Var(\ln(\breve{R}_C)) = \sum_{i=0}^{1} n_{(i)}^{-1} \cdot \dot{e}_{(i)}^T \Sigma_{(i)} \dot{e}_{(i)} \tag{A14}$$

where each component of the vector $\dot{e}_{(i)} = (\dot{e}_{1(i)}, \dot{e}_{2(i)}, \dot{e}_{3(i)}, \dot{e}_{4(i)})^T$ is given by

$\dot{e}_{1(1)} = 0$, $\dot{e}_{2(1)} = \dot{\rho} v_{22(.;1)} \hat{p}_{101} \hat{p}_{000}$, $\dot{e}_{3(1)} = \dot{\rho} v_{33(.;1)} \hat{p}_{10+}^{-1} \hat{p}_{100} (\hat{p}_{110} \hat{p}_{000} - \hat{p}_{010} \hat{p}_{001})$,
$\dot{e}_{4(1)} = -\dot{\rho} \hat{p}_{10+}^{-1} \hat{p}_{100}^{-1} (\hat{p}_{110} \hat{p}_{000} \hat{p}_{101}^2 + \hat{p}_{010} \hat{p}_{001} \hat{p}_{100}^2 + 2 \hat{p}_{110} \hat{p}_{101} \hat{p}_{100} \hat{p}_{000})$;
$\dot{e}_{1(0)} = -v_{11(.;0)} \hat{p}_{01+}^{-1}$, $\dot{e}_{2(0)} = \dot{\rho} v_{22(.;0)} \hat{p}_{01+}^{-1} (\hat{p}_{100} \hat{p}_{011} \hat{p}_{001} - \hat{p}_{000} \hat{p}_{110} \hat{p}_{101})$,
$\dot{e}_{3(0)} = \dot{\rho} v_{33(.;0)} \hat{p}_{00+}^{-1} [\hat{p}_{000} (\hat{p}_{110} \hat{p}_{101} + \hat{p}_{100} \hat{p}_{010}) + 2 \hat{p}_{100} \hat{p}_{010} \hat{p}_{001}]$,
$\dot{e}_{4(0)} = \dot{\rho} v_{44(.;0)} \hat{p}_{00+}^{-1} \hat{p}_{000}^{-1} (\hat{p}_{110} \hat{p}_{101} \hat{p}_{000}^2 - \hat{p}_{100} \hat{p}_{010} \hat{p}_{001}^2)$;
$\dot{\rho} \equiv (\hat{p}_{000} \hat{p}_{110} \hat{p}_{101} + \hat{p}_{100} \hat{p}_{010} \hat{p}_{001})^{-1}$.

Lastly, the variances of the crude and bias-adjusted estimators for Eqs. 7 & 8 are given respectively by

$$\begin{aligned}
&Var(\ln(\hat{\phi}_{E|C})) \approx Var(\ln(\hat{R}_{E|MH}) + Var(\ln(\hat{R}_E)) \ , \\
&Var(\ln(\breve{\phi}_{E|C})) \approx Var(\ln(\breve{R}_{E|MH}) + Var(\ln(\breve{R}_E)) \ , \\
&Var(\ln(\hat{\phi}_{hmg})) = Var(\ln(\hat{R}_{E|C=1}) + Var(\ln(\hat{R}_{E|C=0})) \ , \\
&Var(\ln(\breve{\phi}_{hmg})) = Var(\ln(\breve{R}_{E|C=1}) + Var(\ln(\breve{R}_{E|C=0})) \ .
\end{aligned} \tag{A15}$$

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 211-230.

ISSN 1538 − 9472

Testing the Assumption of Non-differential Misclassification in Case-Control Studies

Tze-San Lee
Western Illinois University
Macomb, IL

Qin Hui
Emory University
Atlanta, GA

One of the not yet solved issues regarding the misclassification in case-control studies is whether the misclassification rates are the same for both cases and controls. Currently, a common practice is to assume that the rates are the same, that is, the non-differential misclassification assumption. However, it has been suspected that this assumption may not be valid in practical applications. Unfortunately, no test is available so far to test the validity of the non-differential misclassification assumption. A method is presented to test the validity of non-differential misclassification assumption in case-control studies with 2 × 2 tables when validation data are not available. First, a theory of exposure operating characteristic curve is developed. Next, two non-parametric methods are presented to test the assumption of non-differential misclassification. Three real-data sets taken from practical applications are used as examples to illustrate the methods.

Keywords: Exposure operating characteristic (EOC) curve, non-differential misclassification, sensitivity, specificity, Youden's index

Introduction

One of the issues regarding the misclassification in case-control studies is whether the misclassification error rates are the same for both cases and controls (Walker & Irwig, 1988). Currently, a common practice is to assume that the rates are the same. This is the so-called "non-differential mis-classification (NDMC)" assumption. Many nice theoretical results are derived under this assumption. For example, in a case-control study with 2×2 contingency table, the adjusted odds ratio is always biased toward the value of the null hypothesis if the misclassification error rates are assumed to be non-differential (Bross, 1954; Goldberg, 1975). However, it is intuitively obvious that the assumption of NDMC might not be valid in many practical applications. Unfortunately, no test is available so far to test the validity of the NDMC assumption.

Dr. Lee is a mathematical statistician at the Centers for Disease Control and Prevention in Chamblee, GA. Email him at: tjl3@cdc.gov. Mr. Hui is an information analyst at the Rollins School of Public Health. Email him at qin.hui@emory.edu.

It is the purpose of this study to propose a test for assessing the validity of NDMC assumption in a case-control study with 2×2 contingency table. First, a theory of exposure operating characteristic curve is developed. Next, two methods are proposed to test the NDMC assumption. Three examples from practical applications are given to illustrate the proposed methods.

Methods

The curve of exposure operating characteristic

The idea of exposure operating characteristic (*EOC*) curve is parallel to that of receiver operating characteristic (*ROC*) curve in medical diagnostic test (Zhou, McClish & Obuchoowski, 2002). Suppose that the collected data for a case-control study is arranged to be given by Table 1. Assume that it is known that Table 1 is possibly misclassified; yet, the truly correct table is unknown. Here the counterfactual thinking comes into playing a crucial role in finding out what the possible true table is, that is, the true table is the counterfactual while the observed misclassified table is the factual (Epstude & Roese, 2008). It may thus be assumed that cell count in the observed table might be over- (or under-) misclassified by a certain number of subjects from the true table. The random variable E in Table 2 is assumed to be correctly classified on the subject's exposure condition, whereas E^* in Table 1 is its misclassified surrogate of E.

Table 1. The observed cell frequencies in a contingency table for a case-control study.

Classified exposure status	*Subject Group*	
	***Y* = 1 (Cases)**	***Y* = 0 (Controls)**
***E** = 1 (exposed)**	n_{11}	n_{10}
***E** = 0 (unexposed)**	n_{01}	n_{00}

Table 2. The [unobserved] true cell frequencies corresponding to Table 1.

Classified exposure status	*Subject Group*	
	***Y* = 1 (Cases)**	***Y* = 0 (Controls)**
***E* = 1 (exposed)**	N_{11}	N_{10}
***E* = 0 (unexposed)**	N_{01}	N_{00}

Let the number of misclassified subjects be given by

$$m_{(i)}^{(j)} = \text{the number of misclassified subjects between true and observed cell frequencies } = N_{ij} - n_{ij} \quad (1)$$

where $m_{(i)}^{(j)}$ (= ±1, ±2, ±3, …) is assumed and N_{ij} can be obtained as $N_{ij} = n_{ij} + m_{(i)}^{(j)}$. It will be clear how to choose the value of $m_{(i)}^{(j)}$ by applying the counterfactual thinking to the observed misclassified cell frequency as shown in the three examples of practical applications later in section 3.

The observed cell frequency (n_{ij}) is said to be under-misclassified if $m_{(i)}^{(j)} > 0$; otherwise it is called over-misclassified. Thus, the sensitivity (*Se*) and specificity (*Sp*) can be calculated for cases and controls as follows:

$$Se(m_{(1)}^{(j)}) = \Pr(E^* = 1 \mid E = 1; D = j) = 1 - \Pr(E^* = 0 \mid E = 1; D = j) = 1 - \frac{|\, m_{(1)}^{(j)} \,|}{N_{1j} + n_{1j}}$$

and (2)

$$Sp(m_{(0)}^{(j)}) = \Pr(E^* = 0 \mid E = 0; D = j) = 1 - \Pr(E^* = 1 \mid E = 0; D = j) = 1 - \frac{|\, m_{(0)}^{(j)} \,|}{N_{0j} + n_{0j}}$$

Note that not all $Se\left(m_{(1)}^{(j)}\right)$ and/or $Sp\left(m_{(0)}^{(j)}\right)$ are feasible. They have to satisfy the following three constraints which are imposed by the cell frequencies in Table 1 (Lee, 2009):

$$Se\left(m_{(1)}^{(j)}\right) + Sp\left(m_{(0)}^{(j)}\right) \neq 1 \tag{3a}$$

$$Se\left(m_{(1)}^{(j)}\right) > \hat{p}_j \tag{3b}$$

$$Sp\left(m_{(0)}^{(j)}\right) > \hat{q}_j \tag{3c}$$

where $\hat{p}_j = n_{1j} / \left(n_{1j} + n_{0j}\right)$ and $\hat{q}_j = 1 - \hat{p}_j$, j = 0, 1.

Varying the values of $m_{(i)}^{(j)}$, it is possible to obtain many feasible sensitivity and specificity pairs. A plot of all feasible pairs of points $\left(Se\left(m_{(1)}^{(j)}\right), 1 - Sp\left(m_{(0)}^{(j)}\right)\right)$ is said to be the *EOC* curve for cases or controls depending on j = 1 or 0. Incidentally, let the number of points on the *EOC* curves for controls and cases be given respectively by m_0 and m_1.

Testing the assumption of non-differential misclassification

In terms of the *EOC* curve, a test on the NDMC is equivalent to following pair of null and alternative hypotheses:

$$H_0 : EOC_1 = EOC_0 \text{ versus } H_1 : EOC_1 \neq EOC_0 \tag{4}$$

There are at least two ways to test equation 4. One way is to use a summary measure, the area under the curve (*AUC*). The measure of *AUC* has been widely used in testing whether the two *ROC* curves associated with the diseased and healthy populations are the same (Hanley & McNeil, 1982). The other way is to use the Kolmogorov-Smirnov test for the bivariate data of sensitivity and specificity pairs.

If a linear interpolation is used to connect all the discrete points on the *EOC* curve, the area under the curve is calculated by using numerical method, namely, the trapezoidal rule. For convenience, let $X(t)$ and $Y(t)$ denote respectively the x-axis (1 – Specificity) and y-axis (Sensitivity), where the variable t represents the misclassified number of subjects. The points $\left(x\left(m_{(0)k}^{(j)}\right), y\left(m_{(1)k}^{(j)}\right)\right)$ lying on the EOC_j curve are given respectively as follows: for $j = 0, 1$; $k = 1, \ldots, m_j$.

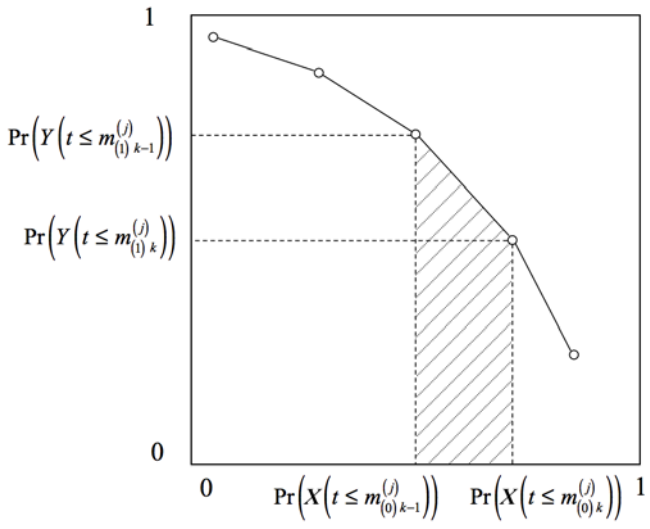


Figure 1. The ordinal dominance graph for the *EOC* curve

$$x\left(m_{(0)k}^{(j)}\right)=P\left(E^{*}=1;t=m_{(0)k}^{(j)}\mid E=0;D=j\right)$$

$$y\left(m_{(1)k}^{(j)}\right)=P\left(E^{*}=1;t=m_{(1)k}^{(j)}\mid E=1;D=j\right)$$

Thus, the EOC_j curve can be viewed as the ordinal dominance (OD) graph with each point on the EOC_j curve with $x\left(m_{(0)k}^{(j)}\right)$ and $y\left(m_{(1)k}^{(j)}\right)$ as its horizontal and vertical coordinates (Fig. 1). Thus, the area under the EOC_j curve (AUC) is calculated by using the trapezoidal rule as follows (Bamber, 1975):

$$\begin{aligned}\theta\equiv AUC&=\sum_{k=1}^{m_j}\frac{1}{2}\left(y\left(m_{(1)k}^{(j)}\right)+y\left(m_{(1)k-1}^{(j)}\right)\right)\cdot\left(x\left(m_{(0)k}^{(j)}\right)-x\left(m_{(0)k-1}^{(i)}\right)\right)\\&=\sum_{k=1}^{m_j}\frac{1}{2}\left(P\left(Y\left(t\le m_{(1)k}^{(j)}\right)\right)+P\left(Y\left(t\le m_{(1)k-1}^{(j)}\right)\right)\cdot P\left(X\left(m_{(0)k}^{(j)}\right)\right)\right)\\&=\sum_{k=1}^{m_j}\left(P\left(Y\left(t\le m_{(1)k-1}^{(j)}\right)\right)+\frac{1}{2}P\left(Y\left(m_{(1)k}^{(j)}\right)\right)\cdot P\left(X\left(m_{(0)k}^{(j)}\right)\right)\right)\\&=P\left(Y\left(t\right)<X\left(t\right)\right)+\frac{1}{2}P\left(Y\left(t\right)=X\left(t\right)\right)\end{aligned}\tag{5}$$

To estimate equation 5, it can be shown that the AUC under the EOC is equivalent to the Wilcoxon-Mann-Whitney test (Pepe, 2003).

Two nonparametric methods for testing equation 4 are thereby summarized as follows:

Method A: The Wilcoxon-Mann-Whitney (WMW) test For each point lying on the EOC_i curve, define the Youden's index (YI) as follows (Zhou, McClish & Obuchowski, 2002):

$$YI\left(P^{(i)}\right)=Se\left(P^{(i)}\right)+Sp\left(P^{(i)}\right)-1,\tag{6}$$

where $P^{(i)}$ is the point lying on the EOC_i curves, i = 0, 1.

Let $P_j^{(1)}$ and $Q_k^{(0)}$ be the points lying on the empirical EOC_i curves for i =1 (cases) and i = 0 (controls) respectively. Define

$$U_{jk} = 1 \text{ if } YI\left(P_j^{(1)}\right) > YI\left(Q_k^{(0)}\right); = 0 \text{ otherwise (assuming no ties).} \quad (7)$$

With equation 5 the null and alternative hypotheses of equation 4 are replaced by

$$H_0' : \theta = 0.5 \text{ versus } H_1' : \theta > 0.5 \quad (8)$$

An unbiased estimator for θ of equation 5 is given by Mee (1990)

$$\hat{\theta} = \frac{1}{m_1 \cdot m_0} \sum_{j=1}^{m_1} \sum_{k=1}^{m_0} U_{jk} \quad (9)$$

where U_{jk} are defined by equation 7 and its variance is given by

$$\text{var}\left(\hat{\theta}\right) = \theta\left(1-\theta\right) / M, \quad (10)$$

where

$$M = m_0 m_1 / \left[\left(m_0 - 1\right)\delta_1 + \left(m_1 - 1\right)\delta_2 + 1\right]$$

and for $\ell = 1, 2$,

$$\delta_\ell = (\theta_\ell - \theta^2) / (\theta(1-\theta)),$$

$$\theta_1 = \Pr(U_{ij} U_{kj} = 1), i \neq k,$$

$$\theta_2 = \Pr(U_{ij} U_{ik} = 1), j \neq k.$$

Note that the estimators for θ_1, θ_2, δ_ℓ, and M are

$$\hat{\theta}_1 = \sum_{i=1}^{m_0} \sum_{j=1}^{m_1} \sum_{k \neq i}^{m_0} U_{ij} U_{kj} / \left[m_0 m_1 \left(m_0 - 1\right)\right]$$

$$\hat{\theta}_2 = \sum_{i=1}^{m_0}\sum_{j=1}^{m_1}\sum_{k\neq j}^{m_1} U_{ij}U_{ik} \Big/ \left[m_0 m_1 (m_1 - 1)\right]$$

$$\tilde{\theta}^2 = \left[m_0 m_1 \hat{\theta}^2 - (m_0 - 1)\hat{\theta}_1 - (m_1 - 1)\hat{\theta}_2 - \hat{\theta}\right] \Big/ \left[(m_0 - 1)(m_1 - 1)\right], \quad (11)$$

$$\tilde{\delta}_\ell = \left(\hat{\theta}_\ell - \tilde{\theta}^2\right) \Big/ \left(\hat{\theta} - \tilde{\theta}^2\right)$$

$$\tilde{M} = m_0 m_1 \Big/ \left[(m_0 - 1)\tilde{\delta}_1 + (m_1 - 1)\tilde{\delta}_2 + 1\right] = \left(\hat{\theta} - \tilde{\theta}^2\right) \Big/ \left(\hat{\theta}^2 - \tilde{\theta}^2\right)$$

Consequently, an estimator of $\text{var}\left(\hat{\theta}\right)$ is given by

$$\hat{\text{var}}\left(\hat{\theta}\right) = \hat{\theta}\left(1 - \hat{\theta}\right) / \tilde{M} \quad (12)$$

Hence, a standard normal z_θ-statistic for testing equation 8 is given by

$$z_\theta = \frac{\hat{\theta} - 0.5}{\sqrt{\hat{\text{var}}\left(\hat{\theta}\right)}} \quad (13)$$

Method B: The Kolmogorov-Smirnov test Let S_{m_1} and T_{m_0} be the sample cumulative distribution function of Youden's index (equation 6) associated with the number of points lying on the *EOC* curves for cases and controls respectively, where S_{m_1} and T_{m_0} are defined respectively as

$$\begin{aligned}
S_{m_1}(t) &= 0, && t < YI\left(P^{(1)}\right)_{(1)}, \\
&= k / m_1 && YI\left(P^{(1)}\right)_{(k)} \le t \le YI\left(P^{(1)}\right)_{(k+1)}, \\
&= 1, && t \ge YI\left(P^{(1)}\right)_{(m1)};
\end{aligned} \quad (14)$$

$$\begin{aligned}
T_{m_0}(t) &= 0, && t < YI\left(Q^{(0)}\right)_{(1)}, \\
&= k / m_0 && YI\left(Q^{(0)}\right)_{(k)} \le t \le YI\left(Q^{(0)}\right)_{(k+1)}, \\
&= 1, && t \ge YI\left(Q^{(0)}\right)_{(m0)},
\end{aligned} \quad (15)$$

where $YI(P^{(1)})_{(i)}$, i = 1,2, …, m_1, and $YI(Q^{(0)})_{(j)}$, j = 1, 2, …, m_0 are the order statistics of Youden's index (equation 6) associated with points lying on the *EOC* curves for cases and controls respectively.

The Kolmogorov-Smirnov test is based on the statistic $K_{m_0m_1}$ defined as

$$K_{m_0m_1} = \sup_t | S_{m_1}(t) - T_{m_0}(t) |, \tag{16}$$

A decision rule for testing equation 4 is given as follows: reject the null hypothesis of equation 4 if the observed number ($K_{m_0m_1}$) (equation 16) is larger than the two-sided critical value K_α, where α is the probability of type I error (Conover, 1971).

Examples

Three examples are used to illustrate how to employ the two methods mentioned in the previous section to test the assumption of non-differential misclassification. The problem now is to calculate the value of sensitivity and specificity when the validation sample data are not available. Here the counterfactual thinking comes into playing the critical role to overcome this barrier (Epstude & Roese, 2008), that is, if only the true (correctly classified) table is known, it is then possible to calculate the value of sensitivity and specificity pair from the observed [misclassified] table by regarding the true table which serves as the "gold standard." Evidently, the potential true table, even though unknown, can be figured out from the observed table as shown below in each of the following three examples. Because it is unknown which potential outcome table is the genuine true table, it is necessary to consider all possible outcome tables figuring out from the observed table as the true table. This leads to a plot of the *EOC* curve separately for the over-/under-misclassification situation in all three examples.

Because the critical values of $K_{0.05}$ are not available for all the following three examples, it was calculated using the large sample approximation $1.36\sqrt{(m_0+m_1)/(m_0m_1)}$, which provided in the last row of Table 17 in (Conover, 1971).

Example 1 The data in Table 3a are taken from a study of deaths caused by landslides that occurred in the State of Chuuk, Federated States of Micronesia, in which a case-control design was used to identify the risk factors (Sanchez et al., 2009). A case was defined to be a person who died as a result of landslides.

Proxies were identified in the surviving villagers to provide information for the decedents, or persons in the control group who were too young to answer questions. Because proxies were used to obtain information on the questions asked in the survey, misclassification was likely to occur. For an illustration, one table was taken from their study regarding whether a person saw natural warning signs (Table 3a). In this example, Exposure = 1 if a person did not see natural warning signs; 0 otherwise.

Assume that the observed table (Table 3a) is misclassified, the potential true table for the over-misclassification situation may be determined by identifying all possible positive integers less than the smallest observed frequency in the (0, 1) cell, $n_{01} = 2$. It turns out there is only one integer which is less than 2. Hence the only potential true (counterfactual) table is given by $N_{11} = 38$, $N_{10} = 26$, $N_{01} = 1$, and $N_{00} = 26$. By using equation 2, $Se_1 = 1 - 1/(38 + 37) = 0.987$ and $Sp_1 = 1 - 1/(1 + 2) = 0.667$; $Se_0 = 1 - 1/(26 + 27) = 0.981$ and $Sp_0 = 1 - 1/(26 + 25) = 0.98$. Hence, the *EOC* curves for cases and controls have just one point $(Se_1, 1 - Sp_1) = (0.987, 0.333)$ and $(Se_0, 1 - Sp_0) = (0.981, 0.02)$ as shown in Table 3b. Although there were 24 true (counterfactual) tables for the under-misclassification situation, only three and seventeen sensitivity and specificity pairs for cases and controls were proved respectively to be feasible, namely, they satisfy all the three constraints of Eqs. 3a-3c. All feasible $(1 - Sp, Se)$ pairs are exhibited as boldface figures in Table 3b. A plot of the *EOC* curves for cases and controls in Example 1 is given in Fig. 2.

Because the results of both methods are not significant (Table 3c), the null hypothesis of equation 4 is not rejected at the significance level of 0.05.

Table 3a. Survey data: whether or not a person saw natural warning signs for cases and controls

	Cases	**Controls**	***Total***
No	37	27	***64***
Yes	2	25	***27***
Total	***39***	***52***	***91***

Table 3b. True (counterfactual) table and the corresponding feasible sensitivity and specificity

Cases				Controls			
Over-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
38	1	0.9867	0.6667	26	26	0.9811	0.9804
Under-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
13	26	0.5200	0.1429	51	1	0.6923	0.0769
14	25	0.5490	0.1481	50	2	0.7013	0.1481
15	24	0.5769	0.1538	49	3	0.7105	0.2143
16	23	0.6038	0.1600	48	4	0.7200	0.2759
18	21	0.6545	0.1739	46	6	0.7397	0.3871
19	20	0.6786	0.1818	45	7	0.7500	0.4375
20	19	0.7018	0.1905	44	8	**0.7606**	**0.4848**
21	18	0.7241	0.2000	43	9	**0.7714**	**0.5294**
22	17	0.7458	0.2105	42	10	**0.7826**	**0.5714**
23	16	0.7667	0.2222	41	11	**0.7941**	**0.6111**
24	15	0.7869	0.2353	40	12	**0.8060**	**0.6486**
25	14	0.8065	0.2500	39	13	**0.8182**	**0.6842**
26	13	0.8254	0.2667	38	14	**0.8308**	**0.7179**
27	12	0.8438	0.2857	37	15	**0.8438**	**0.7500**
28	11	0.8615	0.3077	36	16	**0.8571**	**0.7805**
29	10	0.8788	0.3333	35	17	**0.8710**	**0.8095**
30	9	0.8955	0.3636	34	18	**0.8852**	**0.8372**
31	8	0.9118	0.4000	33	19	**0.9000**	**0.8636**
32	7	0.9275	0.4444	32	20	**0.9153**	**0.8889**
33	6	0.9429	0.5000	31	21	**0.9310**	**0.9130**
34	5	**0.9577**	**0.5714**	30	22	**0.9474**	**0.9362**
35	4	**0.9722**	**0.6667**	29	23	**0.9643**	**0.9583**
36	3	**0.9863**	**0.8000**	28	24	**0.9818**	**0.9796**

Table 3c. Result of applying Methods A and B to Example 1

Method A: Wilcoxon-Mann-Whitney test				**Method B: Kolmogorov-Smirnov test**	
m_0	18	$\tilde{\delta}_1$	-0.05	m_0	18
m_1	4	$\tilde{\delta}_2$	0.17	m_1	4
$\hat{\theta}_1$	0.22	$\tilde{M}$	19	$K_{m_1 m_0}$	0.33
$\hat{\theta}_2$	0.28	$\hat{\theta}$	0.5	$K_{0.05}$	0.75
$\tilde{\theta}^2$	0.24	$\sqrt{\hat{\text{var}}(\hat{\theta})}$	0.11		
		$z_{\hat{\theta}}\sqrt{}$	0		

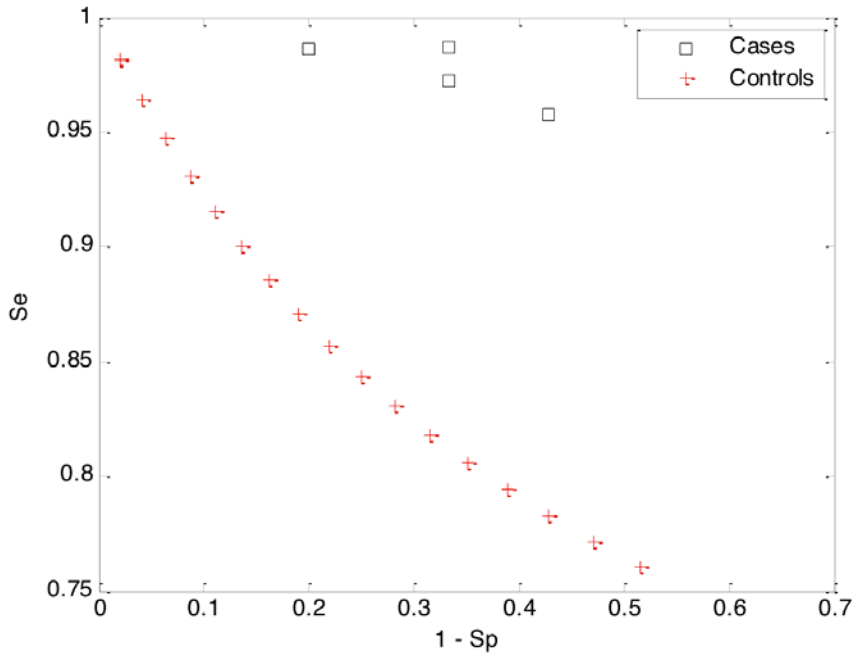


Figure 2. *EOC* curves for cases and controls in Example 1

Example 2 The data in Table 4a are taken from table 7.2 in Schlesselman's book (1982). This data set is in fact a subset of the data from a case-control study of the relation between estrogen use and endometrial cancer in women (Antunes et al., 1979). The use of estrogen is regarded as an exposure risk factor. The rates of exposure for cases and controls are given respectively by $\hat{p}_1$ = 0.3 (= 55/183) and $\hat{p}_0$ = 0.1 (= 19/183). Assume that the exposure data are misclassified for both cases and controls and there is interest in knowing whether their misclassification rates are the same.

By designating the frequency in cell (1, 0) of Table 4a as a free parameter, there were 18 potential true (counterfactual) tables for the over-misclassification scenario. After checking for the feasibility constraints (Eqs. 3b-3c), all 18 pairs of sensitivity and specificity were feasible for cases, while only 17 pairs were feasible for controls. For the under-misclassification scenario, there were 64 potential true (counterfactual) tables. Yet 45 pairs of sensitivity and specificity for cases were feasible, while 30 pairs were feasible for controls. Again, only the top and bottom five pairs are listed in Table 4b. A plot of their *EOC* curves is given in Fig. 3.

Because the results of both methods are not significant (Table 4c), the null hypothesis of equation 4 is not rejected at the significance level of 0.05. By the way, the reason that $\hat{\theta} = 0.42 < 0.5$ is because equation 7 is defined in terms of controls rather than cases, that is, $U_{jk} = 1$ if $YI\left(Q_j^{(0)}\right) > YI\left(P_k^{(1)}\right)$.

Table 4a. Use of oral conjugated estrogen (*OCE*) for endometrial cancer

	Cases	**Controls**	***Total***
User	55	19	***74***
Nonuser	128	164	***292***
Total	***183***	***183***	***366***

Table 4b. True (counterfactual) table and the corresponding feasible sensitivity and specificity

Cases				**Controls**			
Over-misclassification							
N_{11}	**N_{01}**	**Se_1**	**Sp_1**	**N_{10}**	**N_{00}**	**Se_0**	**Sp_0**
54	129	0.9908	0.9961	20	163	0.9744	0.9970

Table 4b Continued

Cases				Controls			
Over-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
53	130	0.9815	0.9922	21	162	0.9500	0.9939
52	131	0.9720	0.9884	22	161	0.9268	0.9908
51	132	0.9622	0.9846	23	160	0.9048	0.9877
50	133	0.9524	0.9808	24	159	0.8837	0.9845
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29	154	0.6905	0.9078	45	138	0.5938	0.9139
28	155	0.6747	0.9046	46	137	0.5846	0.9103
27	156	0.6585	0.9014	47	136	0.5758	0.9067
26	157	0.6420	0.8982	48	135	0.5672	0.9030
25	158	0.6250	0.8951	49	134	0.5588	0.8993
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	169	0.4058	0.8620	60	123	-*	-*
13	170	0.3823	0.8591	61	122	-*	-*
12	171	0.3582	0.8562	62	121	-*	-*
11	172	0.3333	0.8533	63	120	-*	-*
10	173	0.3077	0.8505	64	119	-*	-*
Under-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
73	110	0.8594	0.9244	1	182	-*	-*
72	111	0.8661	0.9289	2	181	0.1905	0.9507
71	112	0.8730	0.9333	3	180	0.2727	0.9535
70	113	0.8800	0.9378	4	179	0.3478	0.9563
69	114	0.8871	0.9421	5	178	0.4167	0.9591
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	123	0.9565	0.9800	14	169	0.8485	0.9850
59	124	0.9649	0.9841	15	168	0.8824	0.9880
58	125	0.9735	0.9881	16	167	0.9143	0.9909
57	126	0.9821	0.9921	17	166	0.9444	0.9939
56	127	0.9910	0.9961	18	165	0.9730	0.9970

*The values of (Se, Sp) are infeasible.

Table 4c. Result of applying the two methods to Example 2

Method A: Wilcoxon-Mann-Whitney test				**Method B: Kolmogorov-Smirnov test**	
m_0	63	$\tilde{\delta}_1$	0.46	m_0	63
m_1	47	$\tilde{\delta}_2$	0.23	m_1	47
$\hat{\theta}_1$	0.23	$\tilde{M}$	82.1	$K_{m_1 m_0}$	0.21
$\hat{\theta}_2$	0.29	$\hat{\theta}$	0.42	$K_{0.05}$	0.26
$\tilde{\theta}^2$	0.17	$\sqrt{\hat{\text{var}}(\hat{\theta})}$	0.05		
		$z_{\hat{\theta}}\sqrt{\quad}$	-1.44		

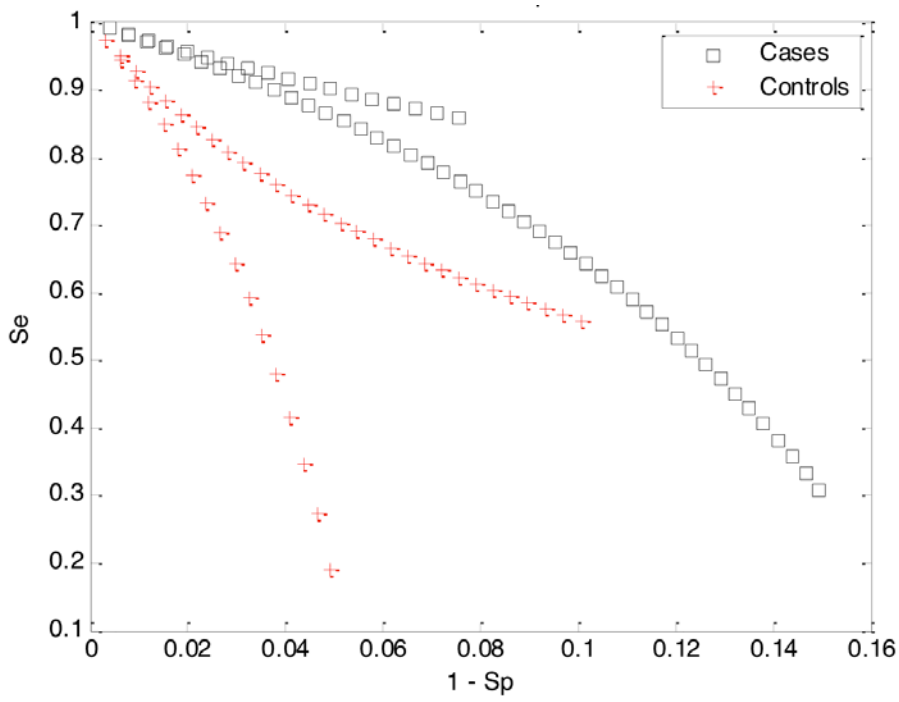


Figure 3. *EOC* curves for cases and controls in Example 2

Example 3 The data of Table 5a are taken from a case control study of sudden infant death syndrome (SIDS) (Greenland, 1988). Among those women who were interviewed, it was asked if she had used the antibiotic medicine during pregnancy. The rate of using the antibiotic medicine for cases and controls were given respectively by $\hat{p}_1$ = 0.22 (= 122/442) and $\hat{p}_0$ = 0.17 (= 101/580). Assume that the interview data are misclassified for both cases and controls and there is interest in knowing whether their misclassification rates are the same.

To do so, it is necessary to obtain their *EOC* curves. To construct the potential true (counterfactual) table, the observed cell frequency n_{10} = 101 were chosen as a reference. For the over-misclassification scenario, the possible values of N_{10} were determined to be integers running from 100 down to 1 (Column 5, Table 5b). After the value of N_{10} was determined, all other cell frequencies were uniquely determined because the column/row totals have to be fixed as the same as that of the observed table. There were 100 potential true (counterfactual) tables. After checking the feasibility constraints imposed by Eqs. 3b-c, all 100 (*Se, Sp*) pairs were feasible for cases, but only 91 (*Se, Sp*) pairs were feasible for controls. To save space, only the top and bottom five pairs are listed (Table 5b). Similarly, for the under-misclassification scenario, the possible values of N_{10} were determined to be integers running from 102 up to 222. There were 121 potential true (counterfactual) tables. Although all 121 true (counterfactual) tables produced feasible pairs of (*Se, Sp*) for controls, only 107 (*Se, Sp*) pairs were feasible for cases. Again, only the top and bottom five pairs are listed (Table 5b). A plot of their *EOC* curves for cases and controls is given respectively in Fig. 4.

Because none of the results obtained from both methods are significant (table 5c), the null hypothesis of equation 4 is not rejected at the significance level of 0.05.

Table 5a. Data of SIDS study of the exposure variable of interview response

	Cases	**Controls**	***Total***
Use	122	101	***223***
No use	442	479	***921***
Total	***564***	***580***	***1144***

Table 5b. True (counterfactual) table and the corresponding feasible sensitivity and specificity

Cases				Controls			
Over-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
123	441	0.9960	0.9989	100	480	0.9950	0.9990
124	440	0.9919	0.9977	99	481	0.9900	0.9979
125	439	0.9879	0.9966	98	482	0.9849	0.9969
126	438	0.9839	0.9955	97	483	0.9798	0.9958
127	447	0.9799	0.9943	96	484	0.9746	0.9948
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
218	346	0.7176	0.8782	14	566	0.2435	0.9167
219	345	0.7155	0.8768	13	567	0.2281	0.9159
220	344	0.7135	0.8754	12	568	0.2124	0.9150
221	343	0.7114	0.8739	11	569	0.1964	0.9141
222	342	0.7093	0.8724	10	570	0.1802	0.9133
Under-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
121	443	0.9959	0.9989	102	478	0.9951	0.9990
120	444	0.9917	0.9977	103	477	0.9902	0.9979
119	445	0.9876	0.9966	104	476	0.9854	0.9969
118	446	0.9833	0.9955	105	475	0.9806	0.9958
117	447	0.9791	0.9944	106	474	0.9758	0.9948
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	545	0.2695	0.8956	204	376	0.6623	0.8795
18	546	0.2571	0.8947	205	375	0.6601	0.8782
17	547	0.2446	0.8938	206	374	0.6580	0.8769
16	548	0.2319	0.8929	207	373	0.6558	0.8756
15	549	0.2190	0.8920	208	372	0.6537	0.8743
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5	559	-*	-*	218	362	0.6332	0.8609
4	560	-*	-*	219	361	0.6313	0.8595
3	561	-*	-*	220	360	0.6293	0.8582
2	562	-*	-*	221	359	0.6273	0.8568
1	563	-*	-*	222	358	0.6254	0.8554

*The values of (Se, Sp) are infeasible.

Table 5c. Result of applying the two methods to Example 3

Method A: Wilcoxon-Mann-Whitney test				**Method B: Kolmogorov-Smirnov test**	
m_0	212	$\tilde{\delta}_1$	0.33	m_0	212
m_1	207	$\tilde{\delta}_2$	0.37	m_1	207
$\hat{\theta}_1$	0.38	$\tilde{M}$	296.9	$K_{m_1 m_0}$	0.05
$\hat{\theta}_2$	0.39	$\hat{\theta}$	0.54	$K_{0.05}$	0.13
$\tilde{\theta}^2$	0.29	$\sqrt{\hat{\text{var}}\left(\hat{\theta}\right)}$	0.03		
		$z_{\hat{\theta}}\sqrt{}$	1.52		

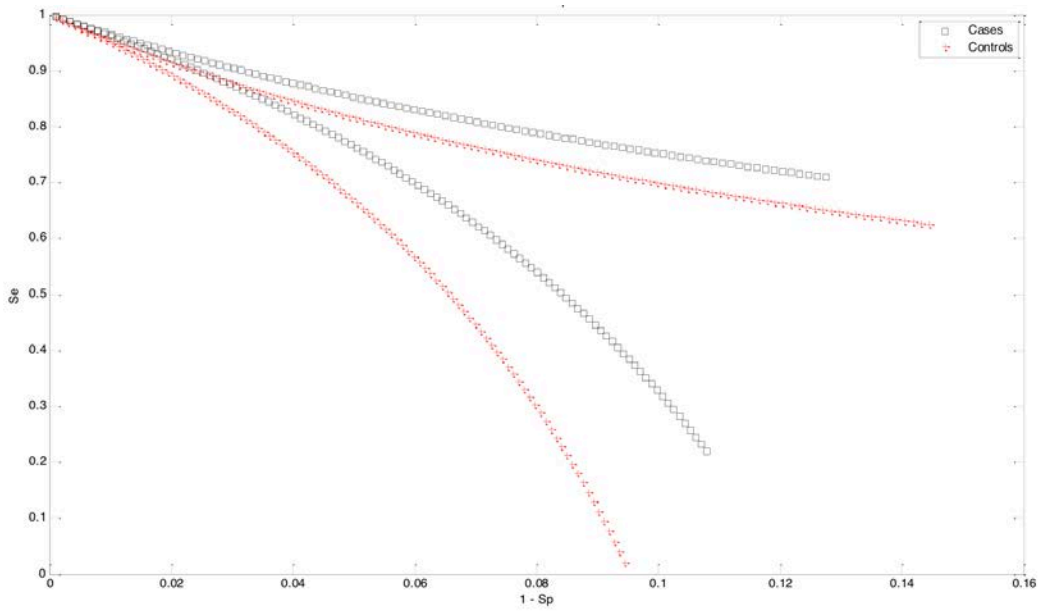


Figure 4. *EOC* curves for cases and controls in Example 3

Discussion

Some comments are worthy to be mentioned below:

1) The *EOC* curve is intrinsically different from that of the *ROC* curve. The *ROC* curve is interested in judging the accuracy of a diagnostic test on the individual's disease status, while the *EOC* curve is concerned with the correct classification of the subject's exposure condition.

2) Unlike the *ROC* curve in which the entire curve is a single continuous curve, the *EOC* curve is comprised of two distinct pieces: one piece of the curve corresponds to the over-misclassification scenario, while the other piece of the curve to the under-misclassification. Further, the *ROC* curve is strictly increasing, whereas the *EOC* curve is monotonically decreasing.

3) It seems that equation 3a is redundant when both equations 3b &3c are satisfied. But, the expression of $Se\left(m_1^{(j)}\right)+Sp\left(m_0^{(j)}\right)-1$ is the determinant of the misclassification matrix for the 2 × 2 contingency table. In fact, equation 3a is the first condition required for the existence of the bias-adjusted proportion estimator (Lee, 2009). Incidentally, the non-singularity of the misclassification matrix is always the first condition required to be satisfied for the existence of the bias-adjusted estimator in other applications too (Lee, 2010, 2011).

4) Method B is preferred to Method A because it is possible that two *EOC* curves are different, but they have the same area.

Conclusion

In this paper a theory of the exposure operating characteristic curve is developed to test the assumption of non-differential misclassification in case-control studies. In terms of the Youden's index two nonparametric methods, the Wilcoxon-Mann-Whitney and Kolmogorov-Smirnov test, are proposed to test whether the two exposure operating characteristic curves are the same for cases and controls. Three real-data examples were used to illustrate the proposed two methods.

Apparently, the idea of the exposure operating characteristic curve for testing the assumption of non-differential misclassification for the 2 × 2 contingency tables presented can be extended to the 2 × *K* or *K* × *K* matched-pair

case-control studies, where $K \geq 3$. This topic will be pursued later in another paper.

Acknowledgements

Part of this paper is based on the result obtained in the Master degree thesis of the second author (QH) (Hui, 2011), which was written under the supervision of the first author (TL). The work in Figure 1 and Tables 3b, 4b and 5b is attributed to QH; the remaining work is attributed to TL. QH is grateful to TL who provided him with an invaluable guidance and constant encouragement while he was writing his thesis.

References

Antunes, C. M. F., Stolley, P. D., Rosenshein, N. B., Davies, J. L., Tonascia, J. A., Brown, C., Burnett, L., Rutledge, A., Pokempner, M., & Garcia, R. (1979). Endometrial cancer and estrogen use: Report of a large case-control study. *New England Journal of Medicine, 300*, 9-13.

Bamber, E. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology, 12*, 387-415.

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics, 10*, 478-486.

Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personnel Social and Psychological Review*, *12*, 168-192.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of American Statistical Association, 70*, 561-567.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine, 7*, 745-757.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (*ROC*) curve. *Radiology, 143*, 29-36.

Hui, Q. (2011). *Testing an Assumption of Nondifferential Misclassification in Case-Control Studies*. Master Degree Thesis, Georgia State University, Atlanta, GA.

Lee, T-S. (2009). Bias-adjusted exposure odds ratio for misclassified data. *The Internet Journal of Epidemiology, 6,* 1-19. http://www.ispub.com/journal/the-internet-journal-of-epidemiology/volume-6-number-2/bias-adjusted-exposure-odds-ratio-for-misclassification-data-1.html.

Lee, T-S. (2010). Misclassified ordinal data in case-control studies. *Journal of Probability and Statistical Science, 8*, 215-226.

Lee, T-S. (2011). Matched-pair studies with misclassified ordinal data. *Journal of Modern Applied Statistical Methods, 10*, 67-76.

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on ageneralization of the Mann-Whitney statistic. *Journal of American Statistical Association, 85*, 793-800.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

Sanchez, C., Lee, T-S., Young, S., Batts, D., Benjamin, J., & Malilay, J. (2009). Risk factors for mortality during 2002 landslides in the State of Chuuk, Federated Staes of Micronesia. *J Disasters, 33*, 705-720.

Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford, England: Oxford University Press.

Walker, S. D., & Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology, 41*, 923-937.

Zhou, X-H., McClish, D. K., & Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. New York: Wiley.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 231-241.

ISSN 1538 − 9472

A Generalized Class of Estimators for Finite Population Variance in Presence of Measurement Errors

Prayas Sharma
Banaras Hindu University
Varanasi, India

Rajesh Singh
Banaras Hindu University
Varanasi, India

The problem of estimating the population variance is presented using auxiliary information in the presence of measurement errors. The estimators in this article use auxiliary information to improve efficiency and assume that measurement error is present both in study and auxiliary variable. A numerical study is carried out to compare the performance of the proposed estimator with other estimators and the variance per unit estimator in the presence of measurement errors.

Keywords: Population mean, study variate, auxiliary variates, mean squared error, measurement errors, efficiency.

Introduction

Over the past several decades, statisticians are paying their attention towards the problem of estimation of parameters in the presence of measurement errors. In survey sampling, the properties of estimators based on data usually presuppose that the observations are the correct measurements on characteristics being studied. However, this assumption is not satisfied in many applications and data is contaminated with measurement errors, such as non-response errors, reporting errors, and computing errors. These measurement errors make the result invalid, which are meant for no measurement error case. If measurement errors are very small and we can neglect it, then the statistical inferences based on observed data continue to remain valid. On the contrary, when they are not appreciably small and negligible, the inferences may not be simply invalid and inaccurate but may often lead to unexpected, undesirable and unfortunate consequences (see Srivastava and Shalabh, 2001). Some important sources of measurement errors in

Prayas Sharma is a Research Fellow in the Department of Statistics. Email him at prayassharma02@gmail.com. Rajesh Singh is Assistant professor in the Department of Statistics. Email at: rsinghstat@gmail.com.

survey data are discussed in Cochran (1968), Shalabh (1997), and Sud and Srivastva (2000). Singh and Karpe (2008, 2010), Kumar et al. (2011a, b) studied some estimators of population mean under measurement error.

Many authors, including Das and Tripathi (1978), Srivastava and Jhajj (1980), Singh and Karpe (2009) and Diana and Giordan (2012), studied the estimation of population Variance σ_y^2 of the study variable y using auxiliary information in the presence of measurement errors. The problem of estimating the population variance and its properties are studied here in the presence of measurement errors.

Consider a finite population U= (U_1, U_2, ... U_N) of N units. Let Y and X be the study variate and auxiliary variate, respectively. Suppose a set of n paired observations are obtained through simple random sampling procedure on two characteristics X and Y. Further assume that x_i and y_i for the i^{th} sampling units are observed with measurement error as opposed to their true values (X_i, Y_i) For a simple random sampling scheme, let (x_i, y_i) be observed values instead of the true values (X_i, Y_i) for i^{th} (i=1.2….n) unit, as

$$u_i = y_i - Y_i \tag{1}$$

$$v_i = x_i - X_i \tag{2}$$

where u_i and v_i are associated measurement errors which are stochastic in nature with mean zero and variances σ_u^2 and σ_v^2, respectively. Further, let the u_i's and v_i's are uncorrelated although X_i's and Y_i's are correlated .

Let the population means of X and Y characteristics be μ_x and μ_y, population variances of (x, y) be (σ_x^2, σ_y^2) and let ρ be the population correlation coefficient between x and y respectively (see Manisha and Singh (2002)).

Notations

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, be the unbiased estimator of population means $\bar{X}$ and $\bar{Y}$, respectively but $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ are not unbiased estimator of (σ_x^2, σ_y^2), respectively. The expected values of s_x^2 and s_y^2 in the presence of measurement error are, given by,

$$E\left(s_x^2\right)=\sigma_x^2+\sigma_v^2$$
$$E\left(s_y^2\right)=\sigma_y^2+\sigma_u^2$$

When the error variance σ_v^2 is known, the unbiased estimator of σ_x^2, is $\hat{\sigma}_x^2=s_x^2-\sigma_v^2>0$, and when σ_u^2 is known, then the unbiased estimator of σ_y^2 is $\hat{\sigma}_y^2=s_y^2-\sigma_u^2>0$.

Define

$$\hat{\sigma}_y^2=\sigma_y^2\left(1+e_0\right)$$
$$\overline{x}=\mu_x\left(1+e_1\right)$$

such that

$$E\left(e_0\right)=E\left(e_1\right)=0,$$
$$E(e_1^2)=\frac{C_x^2}{n}\left(1+\frac{\sigma_v^2}{\sigma_x^2}\right)=\frac{C_x^2}{n\theta_x},$$

and to the first degree of approximation (when finite population correction factor is ignored)

$$E\left(e_0^2\right)=\frac{A_y}{n},\ E\left(e_0e_1\right)=\frac{\lambda C_x}{n}.$$

where,

$$A_y=\left\{\gamma_{2y}+\gamma_{2u}\frac{\sigma_u^4}{\sigma_y^4}+2\left(1+\frac{\sigma_u^2}{\sigma_y^2}\right)^2\right\}\ ,\quad \lambda=\frac{\mu_{12}(x,y)}{\sigma_x\sigma_y^2}\ ,\quad C_x=\frac{\sigma_x}{\mu_x}\ ,\quad \theta_x=\frac{\sigma_x^2}{\sigma_x^2+\sigma_v^2}\ ,$$
$$\theta_y=\frac{\sigma_y^2}{\sigma_y^2+\sigma_u^2}\ ,\ \gamma_{2y}=\beta_2(y)-3\ ,\ \gamma_{2u}=\beta_2(u)-3\ ,\beta_2(u)=\frac{\mu_4(u)}{\mu_2^2(u)},\beta_2(y)=\frac{\mu_4(y)}{\mu_2^2(y)}\ ,$$
$$\mu_4(y)=E(Y_i-\mu_y)^4\ ,\ \mu_4(u)=E(u_i^4).$$

θ_x and θ_y are the reliability ratios of X and Y, respectively, lying between 0 and 1.

Estimator of population variance under measurement error

According to Koyuncu and Kadilar (2010), a regression type estimator t_1 is defined as

$$t_1 = w_1\hat{\sigma}_y^2 + w_2(\mu_x - \bar{x}) \tag{3}$$

where w_1 and w_2 are constants that have no restriction .

Expression (3) can be written as

$$t_1 - \sigma_y^2 = (w_1 - 1)\sigma_y^2 + w_1\sigma_y^2 e_0 - w_2\mu_x e_1 \tag{4}$$

Taking expectation both sides of (4), results in

$$Bias(t_1) = \sigma_y^2(w_1 - 1) \tag{5}$$

Squaring both sides of (4)

$$\left(t_1 - \sigma_y^2\right)^2 = \left[\left(w_1 - 1\right)\sigma_y^2 + w_1\sigma_y^2 e_0 - w_2\mu_x e_1\right] \tag{6}$$

or

$$\begin{aligned}\left(t_1 - \sigma_y^2\right)^2 = \Big[&\sigma_y^4(w_1 - 1)^2 + w_1^2\sigma_y^4 e_0^2 + w_2^2\mu_x^2 e_1^2 + 2(w_1 - 1)w_1\sigma_y^4 e_0 \\ &-2(w_1 - 1)w_2\sigma_y^2\mu_x e_1 - 2w_1w_2\sigma_y^2\mu_x e_0 e_1)\Big]\end{aligned} \tag{7}$$

Simplifying equation (7), taking expectations and using notations, results in the mean square error of t_1 up to first order of approximation, as

$$MSE(t_1) = \left[\sigma_y^4 w_1^2\left(\frac{A_y}{n} + 1\right) + (1 - 2w_1)\sigma_y^4 + w_2^2\mu_x^2\frac{C_x^2}{n\theta_x} - \frac{2w_1w_2\mu_x\sigma_y^2\lambda C_x}{n}\right] \tag{8}$$

In the case, when the measurement error is zero, *MSE* of t_1 without measurement error is given by,

$$MSE^{*}(t_1)=\frac{\sigma_y^4}{n}\left\{\gamma_{2y}+2+n\right\}+(1-2w_1)\sigma_y^4+w_2^2\mu_x^2\frac{C_x^2}{n}-2w_1w_2\mu_x\sigma_y^2\lambda\frac{C_x}{n}\Bigg] \tag{9}$$

and

$$M_{t_1}=\frac{\sigma_y^2}{n}\left[\frac{\sigma_u^4}{\sigma_y^4}\gamma_{2u}+2\left(\frac{\sigma_u^4}{\sigma_y^4}\right)^2+4\frac{\sigma_u^4}{\sigma_y^4}\right]+w_2^2\mu_x^2\frac{C_x^2}{n}\frac{\sigma_v^2}{\sigma_x^2} \tag{10}$$

is the contribution of measurement errors in the *MSE* of estimator t_1.

Differentiating (8) with respect to w_1 and w_2 partially, equating them to zero and after simplification, results in the optimum values of w_1 and w_2, respectively as

$$w_1^{*}=\frac{-\sigma_y^4B}{C^2-AB},\ w_2^{*}=\frac{-\sigma_y^4C}{C^2-AB} \tag{11}$$

where, $A=(\frac{A_y}{n}+1)\sigma_y^4$, $B=\frac{\mu_x^2C_x^2}{n\theta_x}$ and $C=\frac{\sigma_y^2\mu_xC_x\lambda}{n}$.

Using the values of ω_1^{*} and ω_2^{*} from equation (11) into equation (8), gives the minimum *MSE* of the estimator t_2 in terms of *A*, *B* and *C* as

$$MSE(t_1)_{\min}=\left(\frac{\sigma_y^4}{(C^2-AB)}\right)^2\left[\frac{\left(C^2-AB\right)^2}{\sigma_y^4}+3BC^2-AB^2-2BC\right] \tag{12}$$

Another estimator under measurement error

Based on Solanki and Singh (2012), an estimator t_3 is defined as

$$t_2 = \hat{\sigma}_y^2 \left\{ 2\text{-}\left(\frac{\bar{x}}{\mu_x} \right)^{\alpha} \exp\left[\frac{\beta(\bar{x}-\mu_x)}{(\bar{x}+\mu_x)} \right] \right\} \tag{13}$$

where α and β are suitably chosen constants.
Expressing the estimator t_2, in terms of *e*'s is

$$t_2 = \hat{\sigma}_y^2 \left[2-(1+e_1)^{\alpha} \exp\left\{ (\frac{\beta e_1}{2})\left(1+\frac{e_1}{2} \right)^{-1} \right\} \right] \tag{14}$$

Expanding equation (14) and simplifying results in

$$\left(t_2-\sigma_y^2\right) = \sigma_y^2 \left[e_0 - \frac{k}{2}(e_1+e_0e_1) - \frac{e_1^2}{8}\left(k^2-2k\right) \right] \tag{15}$$

where $k = (\beta+2\alpha)$.

On taking expectations of both sides of (15), the bias of the estimator t_3 up to the first order of approximation is obtained as

$$Bias(t_2) = \sigma_y^2 \left[\text{-}\frac{\text{k}}{2}\frac{\lambda C_x}{n} - \left(\frac{k^2-2k}{8} \right)\frac{C_x^2}{n\theta_x} \right] \tag{16}$$

Squaring both sides of (15) and after simplification,

$$\left(t_2-\sigma_y^2\right)^2 = \sigma_y^4 \left[e_0 + \frac{k^2}{4}e_1^2 - ke_0e_1 \right] \tag{17}$$

Taking expectations of (17) and using notations, the *MSE* of estimator t_2 is calculated as

$$MSE(t_2) = \frac{\sigma_y^4}{n\theta_x}\left[A_y\theta_x + \frac{k^2}{4}C_x^2 - k\lambda C_x\theta_x \right] \tag{18}$$

Differentiating equation (18) with respect to *k* and equating to zero and after simplification the optimum value of *k* is

$$k^{*} = 2\frac{\lambda\theta_x}{C_x} \tag{19}$$

Putting the optimum value of k from (19) to (18), results in the minimum *MSE* of estimator t_2 as

$$MSE(t_2)_{\min} = \frac{\sigma_y^4}{n}\left[A_y - \lambda^2\theta_x\right] \tag{20}$$

Remark:

Singh and Karpe (2009) defined a class of estimator for σ_y^2 as

$$t_d = \hat{\sigma}_y^2 d(b) \tag{21}$$

where, $d(b)$ is a function of b such that $d(1)= 1$, and certain other conditions, similar to those given in Srivastava (1971). The minimum *MSE* of t_d is given by,

$$MSE(t_d)_{\min} = \frac{\sigma_y^4}{n}\left[A_y - \lambda^2\theta_x\right] \tag{22}$$

which is the same as the minimum *MSE* of estimator t_2, given in equation (20).

A General Class of Estimators

A general class of estimator t_3 is proposed as

$$t_3 = \left[m_1\hat{\sigma}_y^2 + m_2\left(\mu_x - \bar{x}\right)\right]\left\{2\text{-}\left(\frac{\bar{x}}{\mu_x}\right)^{\alpha}\exp\left[\frac{\beta\left(\bar{x}-\mu_x\right)}{\left(\bar{x}+\mu_x\right)}\right]\right\} \tag{23}$$

Where m_1 and m_2 are constants chosen so as to minimize the mean squared error of the estimator t_3.

Equation (23) can be expressed in terms of *e*'s as

$$t_3 = \left[m_1\sigma_y^2 + m_1\sigma_y^2 e_0 - m_2\mu_x e_1 \right]\left[1 - \frac{k}{2}e_1 - \frac{\left(k^2 - 2k\right)}{8}e_1^2 \right] \tag{24}$$

Expanding equation (24) and subtracting σ_y^2 from both sides, results in

$$\begin{aligned}\left(t_3 - \sigma_y^2\right) = \Bigg[& \left(m_1 - 1\right)\sigma_y^2 - \frac{k}{2}m_1\sigma_y^2 e_1 + m_1\sigma_y^2 e_0 - m_2\mu_x e_1 \\ & - \frac{e_1^2}{8}\sigma_y^2 m_1\left(k^2 - 2k\right) - \frac{\sigma_y^2 m_1 k}{2}e_0 e_1 + \frac{k}{2}m_2\mu_x e_1^2 \Bigg]\end{aligned} \tag{25}$$

On taking expectations of both sides of (25) the bias of the estimator t_3 up to the first order approximation is obtained as

$$Bias\left(t_3\right) = \left(m_1 - 1\right)\sigma_y^2 - \frac{1}{8}\sigma_y^2 m_1\left(k^2 - 2k\right)\frac{C_x^2}{n\theta_x} - \frac{\sigma_y^2 m_1 k}{2}\frac{\lambda C_x}{n} + \frac{k}{2}m_2\mu_x\frac{C_x^2}{n\theta_x} \tag{26}$$

Squaring both sides of (25), results in

$$\left(t_3 - \sigma_y^2\right)^2 = \left[\left(m_1 - 1\right)\sigma_y^2 - \frac{k}{2}m_1\sigma_y^2 e_1 + m_1\sigma_y^2 e_0 - m_2\mu_x e_1 \right]^2 \tag{27}$$

Simplifying equation (27) and taking expectations both sides the *MSE* of estimator t_3 up to the first order of approximation is obtained as

$$MSE(t_3) = \left[\left(1 - 2m_1\right)\sigma_y^4 + m_1^2 P + m_2^2 Q - m_1 m_2 R \right] \tag{28}$$

where $P = \left(1 + \frac{A_y}{n} + \frac{k^2 C_x^2}{4n\theta_x} - \frac{k}{n}\lambda C_x \right)\sigma_y^4$, $Q = \frac{\mu_x^2 C_x^2}{n\theta_x}$ and $R = \sigma_y^2\left(k\frac{C_x^2}{\theta_x} + 2\lambda C_x \right)\frac{\mu_x}{n}$.

Minimizing *MSE* t_3 with respect to m_1 and m_2 the optimum values of m_1 and m_2 is

$$m_1^* = \frac{-4Q\sigma_y^4}{R^2 - 4PQ} \quad \text{and} \quad m_2^* = \frac{-2R\sigma_y^4}{R^2 - 4PQ}$$

Putting the optimum values of m_1 and m_2 in equation (28) results in the minimum *MSE* of estimator t_3 as

$$MSE(t_3) = \sigma_y^4 \left[1 = \frac{4\sigma_y^4 Q}{\left(4PQ - R^2\right)} \right] \tag{29}$$

Empirical Study

Data Statistics:

The data used for empirical study was taken from Gujrati and Sangeetha (2007) - pg, 539., where,

Y_i = True consumption expenditure,
X_i = True income,
y_i = Measured consumption expenditure,
x_i = Measured income.

From the data given we get the following parameter values:

Table 1. Parameter values from empirical data

N	μ_y	μ_x	σ_y^2	σ_x^2	ρ	σ_u^2	σ_v^2
10	127	170	1278	3300	0.964	36.0	36.0

Table 2. Showing the *MSE* of the estimators with and without measurement errors

Estimators	***MSE* without meas. Error**	**Contribution of meas. Errors in *MSE***	***MSE* with meas. Errors**
$\hat{\sigma}_y^2$	245670	35458	281128
t_1	229734	30354	260088

Table 2 continued.

Estimators	*MSE* without meas. Error	Contribution of meas. Errors in *MSE*	*MSE* with meas. Errors
$t_{2\min}$	245411	35461	280872
$t_{3\min}(\alpha=1,\beta=0)$	247440	30442	277862
$(\alpha=0,\beta=1)$	234402	30555	267957
$(\alpha=1,\beta=1)$	268144	30219	298363
$(\alpha=1,\beta=-1)$	234402	33555	267957
$(\alpha=0,\beta=-1)$	231969	30600	262569
$(\alpha=-0.9,\beta=2)$	**229145**	**30365**	**259510**

Conclusion

Table 2 shows that the *MSE* of proposed estimator t_3 (for $\alpha=-0.9,\beta=2$) is minimum among all other estimators considered. It is also observed that the effect due to measurement error on the estimator t_1 and usual estimators is less than the effect on the estimator t_2 under measurement error for this given data set.

References

Allen, J., Singh, H. P., & Smarandache, F. (2003). A family of estimators Of population mean using multi auxiliary information in presence of measurement errors. *International Journal of Social Economics 30*(7), 837–849.

Cochran, W. G. (1968). Errors of Measurement in statistics. *Technometrics 10*, 637-666

Das, A. K., & Tripathi, T. P. (1978). Use of auxiliary information in estimating the Finite population variance. *Sankhya C 4*, 139 - 148

Diana, G., & Giordan, M. (2012). Finite Population Variance Estimation in Presence of Measurement Errors. *Communication in Statistics Theory and Methods, 41*, 4302-4314.

Gujarati, D. N., & Sangeetha (2007). *Basic econometrics*. McGraw – Hill.

Koyuncu, N., & Kadilar, C. (2010). On the family of estimators of Population mean in stratified sampling.Pakistan *Journal of Statistics, 26*, 427-443.

Kumar, M., Singh, R., Singh, A. K., & Smarandache, F. (2011a). Some ratio Type estimators under measurement errors. *World Applied Sciences Journal, 14*(2), 272 - 276.

Kumar, M., Singh, R., Sawan, N., & Chauhan, P. (2011b). Exponential ratio method Of estimators in the presence of measurement errors. *International Journal of Agricultural and Statistical Sciences 7*(2), 457-461.

Manisha, M., & Singh, R. K. (2002). Role of regression estimator involving Measurement errors. Brazilian *Journal of Probability and Statistics 16*, 39- 46.

Shalabh. (1997). Ratio method of estimation in the presence of measurement errors. *Journal of Indian Society of Agricultural Statistics 50*(2), 150– 155.

Singh, H. P. & Karpe, N. (2008). Ratio product estimator for population mean in presence of measurement errors. *Journal of Applied Statistical Sciences, 16*(4), 49-64.

Singh, H. P. & Karpe, N. (2009). Class of estimators using auxiliary Information for estimating finite population variance in presence of measurement errors. *Communication in Statistics Theory and Methods, 38*, 734-741.

Singh, H. P. & Karpe, N. (2010). Effect of measurement errors on the Separate And combined ratio and product estimators in Stratified random sampling. *Journal of Modern Applied Statistical Methods, 9*(2), 338-402.

Solanki R., Singh H.P., & Rathour A. (2012). An alternative estimator for estimating the finite population mean using auxiliary information in sample surveys. *ISRN Probability and Statistics*, doi:10.5402/2012/657682.

Srivastava, M. S. (1971). On Fixed-Width Confidence Bounds for Regression Parameters, *Annals of Mathematical Statistics, 42*, 1403-1411.

Srivastava, A., K., & Shalabh. (2001). Effect of Measurement Errors On the Regression Method of Estimation in Survey Sampling. *Journal of Statistical Research, 35*(2), 35-44.

Srivastava, S. K., & Jhajj, H.S. (1980) A class of estimators using auxiliary information for estimating finite population variance. *Sankhya Ser. C 42*, 87-96.

Sud, U. C., & Srivastava, S. K. (2000). Estimation of population mean in repeat surveys in the presence of measurement errors. *Journal of the Indian Society of Agricultural Statistics, 53*(2), 125-133.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 242-255.

ISSN 1538 − 9472

How Good is Best? Multivariate Case of Ehrenberg-Weisberg Analysis of Residual Errors in Competing Regressions

Stan Lipovetsky
GfK Custom Research North America
Minneapolis, MN

A.S.C. Ehrenberg first noticed and S. Weisberg then formalized a property of pairwise regression to keep its quality almost at the same level of precision while the coefficients of the model could vary over a wide span of values. This paper generalizes the estimates of the percent change in the residual standard deviation to the case of competing multiple regressions. It shows that in contrast to the simple pairwise model, the coefficients of multiple regression can be changed over a wider range of the values including the opposite by signs coefficients. Consideration of these features facilitates better understanding the properties of regression and opens a possibility to modify the obtained regression coefficients into meaningful and interpretable values using additional criteria. Several competing modifications of the linear regression with interpretable coefficients are described and compared in the Ehrenberg-Weisberg approach.

Keywords: Pairwise and multiple regression, residual deviation change, Ehrenberg-Weisberg analysis

Introduction

In a fascinating work by A.S.C. Ehrenberg (1982) it was shown that the coefficients of pairwise regression can be varied over a wide span of values yet the modified model would still have a high quality of fit. Andrew Ehrenberg was a famous English statistician and marketing scientist recognized as the founder of probability models for consumer buying behavior (Ehrenberg, 1959, 1966, 1988; Fader and Hardie, 2009), and a prolific educator in statistics (for several examples, see Ehrenberg, 1981, 1983a,b). As Ehrenberg found, "The residuals from a least squares regression equation are hardly any smaller than those from many other possible lines" (1982, p. 364), and "markedly different equations give almost as good a fit as the least-squares regression equation itself" (1983a, p. 526). The

Dr. Lipovetsky is Senior Research Director at the Research Center for Excellence, GfK CRNA. Email him at: stan.lipovetsky@gfk.com.

technique considered by Ehrenberg was also described by S. Weisberg (1985, p. 68-70) in a convenient analytical form, thus, it will be called the Ehrenberg-Weisberg, or EW analysis.

EW analysis had been developed for pairwise models, but the current paper generalizes it to multiple regression where the results are even more interesting – in particular, it is even possible to change all the predictors' coefficients to the opposite sign, and still have almost the same precision of model fit. Such results show that the coefficients of linear regression can be adjusted by some additional criteria, where the coefficients become meaningful and the quality of the model stays high.

Regression modeling is widely used for statistical analysis and prediction in various problems of applied research. The main tool of regression modeling is the ordinary multiple linear least squares (OLS) regression which yields the best quality of data fit estimated by the minimum residual square error achieved by the aggregate of the predictors. However, OLS was not designed to obtain meaningful coefficients for individual predictors, and it is prone to multicollinearity effects which impact the coefficients' values and directions. Multicollinearity can make confidence intervals so wide that coefficients are incorrectly identified as insignificant, theoretically important variables receive negligible coefficients, or the coefficients have signs opposite to those of the corresponding pair correlations, so it is hardly possible to identify the individual predictors' importance in the regression (Grapentine, 1997; Mason and Perreault, 1991). Multicollinearity makes the covariance matrix of predictors close to singular, so its inversion yields inflated regression coefficients, pushing them to large values of both signs. It is difficult to use such an OLS solution for the analysis of key drivers, either by the coefficients or by the net effects (shares of the coefficient of multiple determination related to the predictors impact).

In the statistical literature and social sciences the effects of multicollinarity are explained by the so-called enhance, synergism, suppression, and masking effects among the predictors (Lipovetsky and Conklin, 2004). But such an explanation hardly helps to the interpretation and analysis of the regression results in applied research. For instance, in customer satisfaction studies in marketing research, the direction of the predictors' influence on the dependent variable is often known in advance. Suppose, the key drivers should all have a positive impact on overall satisfaction and it is evidenced by the pair correlations. But in OLS regression many coefficients turned out to be negative, so it is hardly possible to interpret the model and estimate the individual driver's importance. It is also difficult to use such a model for predicting a lift in the output because it is

not clear whether to increase or decrease a presumably useful variable if it has a negative sign in the model.

This article describes the features of EW analysis and its application to several modifications of multiple regression. One of those is the so-called Shapley value regression which is based on cooperative game theory used for finding the predictors' importance and then adjusting the regression coefficients via a nonlinear optimizing procedure. Another approach uses several modifications of the enhanced ridge regression technique to produce interpretable coefficients with a high overall quality of the model. A nonlinear parameterization of the coefficients of linear regression is also used in several forms to obtain sparse regression models with the features of interest. And finally, a model based on the elasticity criterion applied for building regression coefficients by data gradients is used for a comparison with OLS. In contrast to OLS, all the modified models are meaningful and easily interpretable, and have a quality of fit very close to the maximum defined by the OLS regression (for more detail on these models see (Lipovetsky and Conklin, 2001, 2010 a,b; Lipovetsky, 2009, 2010 a,b).

This paper is organized as follows: the next section describes the characteristics of EW for multiple and pairwise regressions, followed by a description of numerical simulations and a comparison of several modified regression solutions. A summary concludes the paper.

Ehrenberg-Weisberg Analysis

Consider briefly some relations from regression analysis needed for further development. For centered and normalized (by the standard deviations) dependent y_i and n design variables $x_{i1},...,x_{in}$ (i = 1,2,…,N – number of observations), a multiple linear regression model is:

$$y_i = b_1 x_{i1} + b_2 x_{i2} + ... + b_n x_{in} + e_i \tag{1}$$

where e_i denotes deviations from the model, and b are beta-coefficients of the standardized regression. In matrix form (1) can be represented as $y = Xb + e$, where y and e are the vectors of Nth order, and X is the matrix of N by n order. The least-squares objective is:

$$\begin{aligned} S^2 &= \| e \|^2 = \| y - Xb \|^2 = (y - Xb)'(y - Xb) \\ &= y'y - 2b'X'y + b'X'Xb = 1 - 2b'r + b'Rb, \end{aligned} \tag{2}$$

where for the standardized variables it is $y'y = 1$, the vector of pair correlations between y and each of n predictors x is $X'y = r$, and the matrix of pair correlations between the xs is $X'X = C$, and the prime denotes transposition. Minimizing by vector b yields the normal system of equations and its solution

$$Cb = r, \quad b = C^{-1}r, \tag{3}$$

where C^{-1} is inverted correlation matrix. Vector b (3) presents coefficients of the ordinary least squares, or OLS, regression. With OLS estimates b, the minimum residual sum of squares (2) and corresponding to it coefficient of multiple determination R^2 are defined as:

$$S^2 = 1 - b'r, \quad R^2 = 1 - S^2 = b'r = b'Cb \tag{4}$$

where r' is a transposed row-vector of correlations of x-s with y. The coefficient of multiple determination is always non-negative and less than one, its other properties are considered, for instance, in (Reisinger, 1997).

Next, describe EW, or Ehrenberg-Weisberg analysis deriving it from the beginning for the general case of multiple regression. Suppose each jth coefficient of regression b_j is changed with the term k_j, so the modified coefficients are

$$\tilde{b}_j = k_j b_j \tag{5}$$

or in the matrix form $\tilde{b} = diag(k)\,b$, where *diag(k)* is the diagonal matrix of terms k_j, and $\tilde{b}$ is the vector of modified coefficients of regression. With the new parameters $\tilde{b}$ the residual sum of squares (2) becomes:

$$\begin{aligned} \tilde{S}^2 &= \| y - X\tilde{b} \|^2 = (y - X\tilde{b})'(y - X\tilde{b}) \\ &= \left((y - Xb) - X(\tilde{b} - b)\right)' \left((y - Xb) - X(\tilde{b} - b)\right) \\ &= (y - Xb)'(y - Xb) - 2(\tilde{b} - b)'X'(y - Xb) + (\tilde{b} - b)'X'X(\tilde{b} - b) \end{aligned} \tag{6}$$

Taking (3) into account, the middle item in (6) equals zero because $X'(y - Xb) = r - Cb = 0$, so (6) can be reduced to:

$$\tilde{S}^2 = S^2 + (\tilde{b} - b)'C(\tilde{b} - b) = S^2 + b'\left(diag(k) - I\right)C\left(diag(k) - I\right)b \qquad (7)$$

In a simple case when all coefficients are changed by the same term $k_j = k$, taking (4) into account, the expression (7) can be reduced to:

$$\tilde{S}^2 = S^2 + (1-k)^2 b'Cb = S^2 + (1-k)^2 R^2 \qquad (8)$$

Dividing (8) by S^2 and using (4) yields the relation:

$$\frac{\tilde{S}^2}{S^2} = 1 + (1-k)^2 \frac{R^2}{1-R^2} \qquad (9)$$

For the simple pairwise regression by only one predictor ($n = 1$ in (1)) this formula coincides with the one obtained by Ehrenberg and Weisberg up to the change of the multiple correlation R to the pair correlation, $R^2 = r^2$. Taking the square root of (9) produces a quotient of the standard errors of OLS to the modified model expressed as:

$$\left(\frac{\tilde{S}^2}{S^2}\right)^{1/2} = \left(1 + (1-k)^2 \frac{R^2}{1-R^2}\right)^{1/2} \qquad (10)$$

It is the formula given in Weisberg (1985, p. 69) for the pairwise model with $R^2 = r^2$. Because the OLS solution has minimum standard error, (10) can be represented as

$$\left(1 + (1-k)^2 \frac{R^2}{1-R^2}\right)^{1/2} = 1 + d \qquad (11)$$

where $d > 0$ denotes the relative difference of the modified model's and OLS standard errors.

If d is assumed to be at a desirable level, for example, 5% or 10% , then it is possible find the range of k values for which the regression coefficient can be changed but the standard error will be kept within a d% increase from the minimum OLS standard error value:

$$\left(\tilde{S}^2\right)^{1/2} / \left(S^2\right)^{1/2} \le 1+d \tag{12}$$

For this aim, the inequality for k is solved as:

$$\left(1+(1-k)^2 \frac{R^2}{1-R^2}\right)^{1/2} \le 1+d \tag{13}$$

and the solution is:

$$1-\left((2d+d^2)\frac{1-R^2}{R^2}\right)^{1/2} \le k \le 1+\left((2d+d^2)\frac{1-R^2}{R^2}\right)^{1/2} \tag{14}$$

So for regression coefficient change with the term k (5) in the range (14) the inequality (12) is satisfied. With R^2 close to 1 the range (14) is narrow, but with small R^2 the modified coefficient of regression (5) can vary in the wide span without changing much of the residual error. For example, if d = 5%, the span (14) is given by the inequalities:

$$1-0.32\left(\frac{1-R^2}{R^2}\right)^{1/2} \le k \le 1+0.32\left(\frac{1-R^2}{R^2}\right)^{1/2} \tag{15}$$

or the span for the regression coefficient keeping the residual error in the limit of d = 10% is:

$$1-0.46\left(\frac{1-R^2}{R^2}\right)^{1/2} \le k \le 1+0.46\left(\frac{1-R^2}{R^2}\right)^{1/2} \tag{16}$$

It could seem that for small R^2 (for instance if |R|<0.3 in (15), or |R|<.4 in (16)) k can even be negative, so the regression changes its direction. However, for the pairwise regression it is not so, and it is not so for the multiple regression if all parameters of change are constant, $k_j = k$. Indeed, using the coefficients of multiple determination of OLS (4) and of the modified regression $\tilde{R}^2 = 1-\tilde{S}^2$, the equality (8) is represented as:

$$1-\tilde{R}^2 = 1-R^2+(1-k)^2R^2 \tag{17}$$

which reduces to:

$$\tilde{R}^2 = k(2-k)R^2 \tag{18}$$

To keep $\tilde{R}^2 \geq 0$, the values of *k* should belong to the range $0 \leq k \leq 2$. Thus, *k* in (14)-(16) cannot become negative for the pairwise regression, and the same holds for multiple regression in the case where all *k* are equal. The sufficient condition to have $0 \leq k \leq 2$, so $\tilde{R}^2 \geq 0$, is to keep the square roots in (14) less than one:

$$\left((2d+d^2)\frac{1-R^2}{R^2}\right)^{1/2} \leq 1 \tag{19}$$

which can be represented more concisely as follows:

$$(1+d)^2(1-R^2) \leq 1 \tag{20}$$

Thus, for a given value of R^2 the percent *d* satisfying the condition (20) which guarantees the modified $\tilde{R}^2$ to be positive should be chosen.

Continuing with EW description for multiple regression in a general case where different parameters of change are assigned to each coefficient, similar to the transformation of (8) to (9), the general expression (7) can be presented in explicit form as:

$$\frac{\tilde{S}^2}{S^2} = 1+\frac{1}{1-R^2}\left(\sum_{j=1}^{n}(1-k_j)^2b_j^2 + 2\sum_{j>q}^{n}(1-k_j)(1-k_q)b_jb_qr_{jq}\right) \tag{21}$$

where r_{jq} are the pair correlations between x_j and x_q. The terms with $1 - k_j$ in (21) modify the inputs from b_j^2 (the so-called pure net-effects of each predictor) and from $b_jb_qr_{jq}$ (the so-called mixed net-effects of the predictors) into the coefficient of multiple determination. If only one coefficient of regression is changed, say, $k_j \neq 1$, and all the others are kept intact ($k = 1$) then the ratio (21) reduces to:

$$\frac{\tilde{S}^2}{S^2} = 1 + (1-k_j)^2 \frac{b_j^2}{1-R^2} \tag{22}$$

From (22) with the net effect of the *j*th predictor in the numerator, it is easy to derive the relations (10)-(14) for considering a model with only one modified coefficient. But a general case of different changes for all the coefficients (21) can be studied in numerical simulations.

Numerical Simulation and Examples

Consider the case of two predictors, $n = 2$, trying several values of pairwise correlations r_{y1} and r_{y2} of y with x_1 and x_2, and the r_{12} correlation between two predictors taken within the allowed range of the values:

$$r_{y1}r_{y2} - \sqrt{(1-r_{y1}^2)(1-r_{y2}^2)} \;\le\; r_{12} \;\le\; r_{y1}r_{y2} + \sqrt{(1-r_{y1}^2)(1-r_{y2}^2)} \tag{23}$$

Table 1. Numerical simulation with various *k*: pair correlations, OLS regressions, *k*-terms, modified regressions, and residual STD change.

Pair correlations			OLS regression			Terms of change		Modified regression			STD change
r_{y1}	r_{y2}	r_{12}	b_1	b_2	R^2	k_1	k_2	$\tilde{b}_1$	$\tilde{b}_2$	$\tilde{R}^2$	d
-0.75	0.75	-0.900	-0.395	0.395	0.592	-0.1	2.0	0.039	0.789	0.556	0.043
-0.75	0.75	-0.900	-0.395	0.395	0.592	0.1	2.0	-0.039	0.789	0.562	0.036
-0.75	0.75	-0.900	-0.395	0.395	0.592	0.5	2.0	-0.197	0.789	0.538	0.065
-0.75	0.75	-0.813	-0.414	0.414	0.621	2.0	-0.1	-0.828	-0.041	0.548	0.091
-0.75	0.75	-0.813	-0.414	0.414	0.621	2.0	0.1	-0.828	0.041	0.561	0.076
-0.75	0.75	-0.813	-0.414	0.414	0.621	2.0	0.5	-0.828	0.207	0.546	0.094
-0.50	0.50	-0.150	-0.435	0.435	0.435	0.5	0.5	-0.217	0.217	0.326	0.092
0.10	0.50	0.739	-0.595	0.940	0.410	0.5	0.5	-0.297	0.470	0.308	0.084
0.50	0.50	0.100	0.455	0.455	0.455	0.5	0.5	0.227	0.227	0.341	0.099
0.50	0.75	0.490	0.175	0.664	0.586	2.0	0.5	0.349	0.332	0.502	0.097
0.50	0.75	0.604	0.074	0.705	0.566	5.0	0.5	0.369	0.353	0.480	0.094
0.50	0.75	0.719	-0.081	0.808	0.566	-2.0	0.5	0.161	0.404	0.484	0.090
0.75	0.75	0.825	0.411	0.411	0.616	2.0	-0.1	0.822	-0.041	0.550	0.083
0.75	0.75	0.825	0.411	0.411	0.616	2.0	0.1	0.822	0.041	0.562	0.069
0.75	0.75	0.825	0.411	0.411	0.616	2.0	0.5	0.822	0.205	0.545	0.090

Table 1 presents the sets of these three correlations in the first three columns, and in the next three columns there are OLS beta-coefficients of regression (1) and the coefficient of multiple determination R^2 (4). The terms k_j for the modified coefficients are given in the two middle columns of Table 1, then there are the modified coefficients themselves (5), and the corresponding modified coefficient $\tilde{R}^2 = 1 - \tilde{S}^2$ of multiple determination. The last column of Table 1 presents the relative change of the residual standard deviation (STD), which can be expressed via (12) and (21) as follows:

$$d = \sqrt{1 + \frac{(1-k_1)^2 b_1^2 + (1-k_2)^2 b_2^2 + 2(1-k_1)(1-k_2) b_1 b_2 r_{12}}{1-R^2}} - 1 \qquad (24)$$

As shown in Table 1, the change of coefficients can be very noticeable but the change in STD is below 10% of the precision in all fifteen examples given in rows.

Table 2. Numerical simulation with negative *k*: pair correlations, OLS regressions, *k*-terms, modified regressions, and residual STD change.

Pair correlations			OLS regression			Terms of change		Modified regression			STD change
r_{y1}	r_{y2}	r_{12}	b_1	b_2	R^2	$k_1<0$	$k_2<0$	$\tilde{b}_1$	$\tilde{b}_2$	$\tilde{R}^2$	d
-0.50	-0.25	0.628	-0.566	0.106	0.257	-0.1	-2.0	0.057	-0.212	0.016	0.151
-0.50	-0.25	0.628	-0.566	0.106	0.257	-0.1	-2.0	0.028	-0.212	0.039	0.137
-0.50	-0.25	0.628	-0.566	0.106	0.257	-0.1	-1.0	0.028	-0.106	0.016	0.150
-0.50	-0.25	0.796	-0.821	0.403	0.310	-0.1	-1.0	0.082	-0.403	0.003	0.202
-0.50	-0.25	0.796	-0.821	0.403	0.310	-0.1	-1.0	0.041	-0.403	0.023	0.190
-0.50	-0.25	0.796	-0.821	0.403	0.310	-0.1	-0.5	0.041	-0.202	0.031	0.185
-0.25	-0.10	0.603	-0.298	0.080	0.067	-0.1	-2.0	0.015	-0.160	0.002	0.034
-0.25	-0.10	0.603	-0.298	0.080	0.067	-0.1	-1.0	0.015	-0.080	0.003	0.033
-0.25	-0.10	0.796	-0.465	0.270	0.089	-0.1	-0.5	0.023	-0.135	0.002	0.047
0.50	0.75	0.719	-0.081	0.808	0.566	-5.0	-0.1	0.404	-0.040	0.202	0.356
0.50	0.75	0.719	-0.081	0.808	0.566	-2.0	-0.1	0.161	-0.081	0.026	0.497
0.50	0.75	0.719	-0.081	0.808	0.566	-2.0	-0.1	0.161	-0.040	0.082	0.453
0.50	0.75	0.833	-0.409	1.091	0.614	-2.0	-0.1	0.817	-0.055	0.139	0.493
0.50	0.75	0.833	-0.409	1.091	0.614	-1.0	-0.1	0.409	-0.109	0.140	0.491
0.50	0.75	0.833	-0.409	1.091	0.614	-1.0	-0.1	0.409	-0.055	0.194	0.444

Table 2 is organized as Table 1 but it contains both the *k*-terms of negative sign, $k_1 < 0$ and $k_2 < 0$, so the direction of *y*'s connections with the predictors is flipped. Table 2 shows that although the direction of the model coefficients can be changed, the quality of such models is not high, and the precision of STD change could be low too. As can be expected, a model could receive the opposite signs of the coefficients and keep about the same quality of fit mostly in the cases of weak statistical relationships similar to those considered in (Langford et al., 2001).

As it was discussed in the introduction, because of the effects of multicollinearity the coefficients of regression can be found in a wide range of the values of both signs. It can be shown in a simple example of the model with two predictors where the beta-coefficients of regression are defined as follows:

$$b_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}, \quad b_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \tag{25}$$

Suppose all correlations are positive, and x_1 is strongly correlated with x_2, so r_{12} is close to 1. Then the numerators in the coefficients (25) become close to $r_{y1} - r_{y2}$ and $r_{y1} - r_{y2}$, respectively, so of opposite signs. At the same time the denominator $1 - r_{12}^2$ is close to zero, so b_1 and b_2 become big by the absolute value and of opposite signs. It is effect of inflation under multicollinearity, and changing directions of the connection from positive pairwise to opposite by sign in multiple regression. Using various methods of regularization, mentioned in the introduction, meaningful regression coefficients can be obtained. And the EW relative change of the residual standard deviations can be used for comparison of the several competing regression models and checking how far are the residual errors from their OLS minimum value.

Consider a numerical example where several regressions were tried by the data on various cars' characteristics given in (Chambers and Hastie, 1992; and also available in *S-PLUS'2000*, 1999, as "car.all" data). The data describes dimensions and mechanical specifications supplied by the manufacturers and measured by Consumer Reports. The variables are: *y* – Price of a car, US$K; x_1 – Weight, pounds; x_2 – Length overall, inches; x_3 – Wheel base length, inches; x_4 – Width, inches; x_5 – Front Leg Room maximum, inches; x_6 – Front Shoulder room, inches; x_7 – Turning circle radius, feet; x_8 – Displacement of the engine, cubic inches; x_9 – HP, the net horsepower; x_{10} – Tank fuel refill capacity, gallons. The cars' price is estimated in the regression model by the dimensions and specifications variables.

Table 3. Cars example: correlations and several regressions.

Name	variable	r_{yx}	OLS	SV	Grad	RR	RE2	RE3	Exp	Multin
Weight	x_1	0.653	0.278	0.129	0.101	0.053	0.105	0.116	0.088	0.000
Length	x_2	0.533	0.225	0.072	0.083	0.039	0.056	0.062	0.066	0.099
Wheel .base	x_3	0.496	-0.085	0.043	0.077	0.032	0.034	0.038	0.000	0.000
Width	x_4	0.478	-0.144	0.047	0.074	0.029	0.024	0.026	0.000	0.000
Frt.Leg .Room	x_5	0.567	0.245	0.140	0.088	0.063	0.129	0.143	0.258	0.248
Frt.Shld	x_6	0.371	-0.060	0.012	0.057	0.017	0.006	0.007	0.000	0.000
Turning	x_7	0.378	-0.199	0.022	0.059	0.017	0.003	0.003	0.000	0.000
Disp.	x_8	0.642	0.101	0.110	0.100	0.053	0.097	0.107	0.000	0.000
HP	x_9	0.783	0.409	0.191	0.121	0.082	0.293	0.323	0.512	0.528
Tank	x_{10}	0.657	0.160	0.114	0.102	0.056	0.116	0.128	0.085	0.125
	R^2		0.722	0.596	0.503	0.409	0.637	0.645	0.695	0.694
	d			0.205	0.337	0.458	0.143	0.130	0.047	0.049

Table 3 in the first and second numerical columns presents the pair correlations r_{yx} of y with x, and the OLS beta-coefficients (1). All correlations are positive, but four of the ten variables have negative coefficients in the multiple OLS regression, although it has a good coefficient of multiple determination $R^2 = 0.722$. The next seven columns in Table 3 present several modified solutions referred to in the introduction: *SV* – Shapley value model, *Grad* – the model constructed by the data gradients; *RR* – the regular ridge regression, *RE2* and *RE3* – two kinds of the ridge enhanced models, *Exp* and *Multin* – the model with exponential and multinomial-logit parameterization of the coefficients of multiple linear regression. Below each model, its coefficient of multiple determination is shown, together with the EW relative change characteristic of the residual standard deviation *d*.

All the modified models have non-negative coefficients of regression, and their coefficients R^2 are slightly less than the maximum R^2 of OLS. But the more sensitive characteristic of *d* indicates rather clearly that *RR* and *Grad* models are fair, the *SV* and both *RE* models are good, and the *Exp* and *Mult* models give the best variants with less than 5% of the difference in standard deviations. As had been shown with more detail in (Lipovetsky, 2009, 2010a,b), the enhanced and adjusted ridge models systematically outperform regular ridge regression, and

special parameterization techniques produce nonnegative coefficients with a clear, sparse structure in the two last approaches. As an additional useful feature of the *Mult* model, the total of the beta-coefficients equals exactly one, so the coefficients equivalent to the shares of the predictors' impact on the dependent variable. However, if it is desirable to keep and compare all the variables in the model then the *SV* and *ridge* regressions should be used, and the *Grad* model is preferable for an express analysis when no special software is available.

Summary

A modified least squares regression can have better interpretable coefficients and practically the same quality of fit, which can be estimated by the characteristic of the relative change in the residual standard deviation. This paper develops the Ehrenberg-Weisberg estimation of the characteristic of relative change in the residual standard deviation for pair regression to the general case of multiple regression. It shows that the coefficients of ordinary least-squares can be changed over a wide range of values, including the opposite sign, and the quality of fit can still be at an acceptable level. This estimation is applied for a comparison of several regressions with the ordinary least squares model, to identify the modified regressions with interpretable coefficients and good quality of fit. The obtained results help provide a better understanding of the properties of multiple regression, and are useful for theoretical consideration and practical applications of regression modeling and analysis.

References

Chambers, J.M., and Hastie, T.J. (1992). *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks.

Ehrenberg, A.S.C. (1959). The Pattern of Consumer Purchases. *Applied Statistics*, *8*, 26–41.

Ehrenberg, A.S.C. (1966). Laws in Marketing: A Tail-Piece. *Journal of the Royal Statistical Society, Series C, 15*, 257-267.

Ehrenberg, A.S.C. (1981). The Problem of Numeracy. *The American Statistician, 35*, 67-71.

Ehrenberg, A.S.C. (1982). How Good is Best? *Journal of the Royal Statistical Society, Series A, 145*, 364-366.

Ehrenberg, A.S.C. (1983a). Lawlike relationships. In: *Encyclopedia of Statistical Sciences* (N. L. Johnson & S. Kotz, Eds.), *4*, 523-528, NY: Wiley.

Ehrenberg, A.S.C. (1983b). Deriving the Least Squares Regression Equation. *The American Statistician*, *37*, 232.

Ehrenberg, A. S. C. (1988). *Repeat-Buying*. 2nd ed. London: Griffin.

Fader, P.S., and Hardie, B.G.S. (2009). Probability Models for Customer-Base Analysis. *J. of Interactive Marketing, 23*, 61-69.

Grapentine, T. (1997). Managing Multicollinearity. *Marketing Research*, *9,* 11-21.

Langford, E., Schwertman, N., and Owens, M. (2001). Is the Property of Being Positively Correlated Transitive?. *The American Statistician, 55,* 322-325.

Lipovetsky, S. (2009). Linear Regression with Special Coefficient Features Attained via Parameterization in Exponential, Logistic, and Multinomial-Logit Forms. *Mathematical and Computer Modelling*, *49*, 1427-1435.

Lipovetsky, S. (2010a). Enhanced Ridge Regressions. *Mathematical and Computer Modelling, 51*, 338-348.

Lipovetsky, S. (2010b). Meaningful Regression Coefficients Built by Data Gradients. *Advances in Adaptive Data Analysis*, *2*, 451–462.

Lipovetsky, S., and Conklin, M. (2001). Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry*, *17,* 319-330.

Lipovetsky, S., and Conklin, M. (2004). Enhance-Synergism and Suppression Effects in Multiple Regression. *International Journal of Mathematical Education in Science and Technology, 35*, 391-402.

Lipovetsky, S., and Conklin, M. (2010a). Reply to the paper 'Do not adjust coefficients in Shapley value regression'. *Applied Stochastic Models in Business and Industry*, *26*, 203-204.

Lipovetsky, S., and Conklin, M. (2010b). Meaningful Regression Analysis in Adjusted Coefficients Shapley Value Model. *Model Assisted Statistics and Applications, 5*, 251-264.

Mason, C.H., and Perreault, W.D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, *28,* 268-280.

Reisinger, H. (1997) The impact of research designs on R^2 in linear regression models: an exploratory meta-analysis. *Journal of Empirical Generalisations in Marketing Science, 2*, 1-12.

S-PLUS'2000 (1999). Seattle, WA: MathSoft.
Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. New York: Wiley.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 256-268.

ISSN 1538 − 9472

Comparison of Three Calculation Methods for a Bayesian Inference of $P(\pi_1 > \pi_2)$

Yohei Kawasaki
National Center for Global Health and Medicine
Tokyo, Japan

Asanao Shimokawa
Tokyo University of Science
Tokyo, Japan

Etsuo Miyaoka
Tokyo University of Science
Tokyo, Japan

In Bayesian inference, some researchers have examined the difference of binominal proportions using $\theta = \mathrm{P}\left(\pi_1 > \pi_2 - \Delta_0 | X_1, X_2\right)$, where X_i denote binomial random variable with parameter π_i. An approximate method and the MCMC method are compared with an exact method for θ, and results of actual clinical trials using θ are presented.

Keywords: Binominal proportions; Bayesian inference; MCMC method; hypergeometric series.

Introduction

Statistical inference concerning the difference between two independent binominal proportions is often discussed from the frequency rather than the Bayesian viewpoint. Some researchers have examined significant differences in binominal proportions using the index, $\theta = \mathrm{P}\left(\pi_1 > \pi_2 - \Delta_0 | X_1, X_2\right)$, which indicates the difference in the posterior density for two independent binomial proportions that are assumed to be random variables.

Originally, this index can be shown in the framework of frequency theory to be, $\mathrm{P}(Y_1 > Y_2)$, where Y_1 and Y_2 are random variables. The inference for $\mathrm{P}(Y_1 > Y_2)$ can be observed in various fields. In engineering, it is used in the `stress strength model' to evaluate the reliability of an industrial component (see for instance Kotz, et al. (2003)). In clinical research, it is used as an index for the comparison of two groups given different treatments. In addition, this probability corresponds to the area under the receiver operating characteristic (ROC) curve.

Dr. Kawasaki is a Senior Biostatistician at the National Center for Global Health and Medicine. Email at yk_sep10@yahoo.co.jp. Asanao Shimokawa is a graduate student. Dr. Miyaoka is a professor in the mathematics department.

In medicine, it is used as an index for evaluating the validity of a diagnostic method. Indeed, innumerable studies have been conducted for $\mathrm{P}(Y_1 > Y_2)$ in the framework of frequency theory (See for instance Sen (1960, 1967)). As for research papers on this index, Shirahata (1993), Zhou (2008) and Kawasaki and Miyaoka (2010) have published actively in recent years.

Conversely, there have been a number of studies to apply a construction of $\mathrm{P}(Y_1 > Y_2)$ to the Bayesian framework. Basu (1996) concisely showed the use of the Bayesian approach with respect to hypothesis testing. Berry (1995) using superior binomial proportions, presented a detailed comparison between two binomial proportions assumed to be random variables and presented some interesting examples. Zaslavsky (2009, 2010) applied θ to a one-side hypothesis based on a one-sample situation. Kawasaki and Miyaoka (2012) showed an exact expression for θ, and applied θ to a one-side hypothesis based on a two-sample situation.

There are some pending issues with the above-mentioned method. An approximate method and exact method of θ were adopted only while using a conjugate prior. The drawback of the approximate method is that it occasionally leads to a rough result in a small sample. The drawback of the exact method is that it is slightly complicated. In addition, the exact method requires extensive computing time with a large sample size. Hence, a Markov Chain Monte Carlo (MCMC) method is proposed for θ as a solution to these problems.

Methodology

Let X_1 and X_2 denote binomial random variables for n_1 and n_2 trials with parameters π_1 and π_2, respectively. The conjugate prior density for π_i is a beta distribution with parameters α_i and β_i, where $\alpha_i > 0$, $\beta_i > 0$, and $i = 1, 2$. The proposed posterior density for π_i is

$$g_i\left(\pi_i | X_i\right) = \frac{1}{\mathrm{B}\left(a_i, b_i\right)} \pi_i^{a_i - 1} \left(1 - \pi_i\right)^{b_i - 1} \dagger, \tag{1}$$

where $a_i = \alpha_i + x_i$, $b_i = n_i - x_i + \beta_i$, and $\mathrm{B}\left(a, b\right)$ is the proposed beta function. Let $\pi_{i,\dagger post}$ denote the binomial proportion following the posterior density.

Approximate method for θ

θ can be calculated via an approximation using the standard normal table. Assume that a_i and b_i of the posterior density are large. It is necessary to determine a Z-test statistic. The expected difference in the posterior density and the variance in this difference can be expressed as:

$$\mathrm{E}\left(\pi_{1,post} - \pi_{2,post}\right) = \mu_{1,post} - \mu_{2,post}\dagger, \tag{2}$$

$$\mathrm{V}\left(\pi_{1,post} - \pi_{2,post}\right) = \frac{\mu_{1,post}\left(1-\mu_{1,post}\right)}{a_1+b_1+1} + \frac{\mu_{2,post}\left(1-\mu_{2,post}\right)}{a_2+b_2+1}\dagger, \tag{3}$$

where $\mu_{i,post} = a_i / \left(a_i + b_i\right)$ denotes the posterior mean of π_i. The Z_g-test statistic,

$$Z_g = \frac{\left(\pi_{1,post} - \pi_{2,post}\right) - \mathrm{E}\left(\pi_{1,post} - \pi_{2,post}\right)}{\sqrt{\mathrm{V}\left(\pi_{1,post} - \pi_{2,post}\right)}}\dagger, \tag{4}$$

is approximately distributed as the standard normal distribution. Therefore, the approximate probability of θ is given by

$$\begin{aligned}\theta &= \mathrm{P}\left(\pi_1 > \pi_2 | X_1, X_2\right) \\ &\approx 1 - \Phi\left(\frac{-\left(\mu_{1,post} - \mu_{2,post}\right)}{\sqrt{\frac{\mu_{1,post}\left(1-\mu_{1,post}\right)}{a_1+b_1+1} + \frac{\mu_{2,post}\left(1-\mu_{2,post}\right)}{a_2+b_2+1}}}\right)\end{aligned} \tag{5}$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function (CDF) of the standard normal distribution. Thus, the approximate probability can easily be calculated.

Exact method for θ

Kawasaki and Miyaoka (2012) derived the exact expression for θ using the posterior density. The exact expression for θ is

$$\begin{aligned}\theta &= \mathrm{P}\left(\pi_1 > \pi_2 | X_1, X_2\right) \\ &= \frac{\mathrm{B}\left(a_1 + a_2, b_1\right)}{a_2 \mathrm{B}\left(a_1, b_1\right) \mathrm{B}\left(a_2, b_2\right)} \dagger_3 \mathrm{F}_2\left(a_2, 1-b_2, a_1+a_2; 1+a_2, \dagger a_1 + a_2 + b_1; 1\right),\end{aligned} \tag{6}$$

where

$$\dagger_3 \mathrm{F}_2\left(k_1, k_2, k_3; l_1, l_2; 1\right) = \sum_{t=0}^{\infty} \frac{\left(k_1\right)_t \left(k_2\right)_t \left(k_3\right)_t}{\left(l_1\right)_t \left(l_2\right)_t} \frac{1}{t!}, \dagger\dagger k_1 + k_2 + k_3 < l_1 + l_2 \tag{7}$$

is the hypergeometric series, and $\left(k\right)_t$ is the Pochhammer symbol.

MCMC method for θ

A computational procedure for θ using the MCMC method is now introduced. The MCMC method is a means of sampling from a posterior density. A random-walk Metropolis-Hasting algorithm was used as the MCMC Method. Given that the samples come from two independent populations, the posterior joint distribution of π_1 and π_2 is a product of its marginal distributions. For this reason, one can obtain samples from the posterior distribution of $\pi_1 - \pi_2$ by simulating k values from the posterior distribution of π_1 and π_2 using MCMC procedure of SAS, e.g., $\pi_{1,post}^1, \pi_{1,post}^2, \cdots, \pi_{1,post}^k$ and $\pi_{2,post}^1, \pi_{2,post}^2, \cdots, \pi_{2,post}^k \dagger$, respectively. Then, by computing $\pi_{1,post}^1 - \pi_{2,post}^1, \dagger \pi_{1,post}^2 - \pi_{2,post}^2 \dagger, \cdots, \pi_{1,post}^k - \pi_{2,post}^k \dagger$, the simulated values from the posterior distribution of $\pi_{1,post} - \pi_{2,post}$ are obtained. The posterior samples obtained by the MCMC method after the burn-in period are $\delta_1, \delta_2, \cdots, \delta_k$. Let $\Delta_1, \Delta_2, \cdots, \Delta_k$ be independent identically distributed random variables with distribution function *F*. The posterior samples is the observed value of $\Delta_1, \Delta_2, \cdots, \Delta_k$. Note the fact that $\theta = \mathrm{P}(\pi_{1,post} > \pi_{2,post})$ equals $\theta = \mathrm{P}(\pi_{1,post} - \pi_{2,post} > 0)$. Thus, θ can be expressed as,

$$\begin{aligned}\theta &= \mathrm{P}\left(\pi_1 > \pi_2 | X_1, X_2\right) \\ &= \mathrm{P}\left(\pi_1 - \pi_2 > 0 | X_1, X_2\right) \approx 1 - \hat{\mathrm{F}}_k\left(0\right)\end{aligned} \tag{8}$$

where

$$\hat{\mathrm{F}}_{\mathrm{k}}(\mathrm{s}) = \frac{1}{k}\sum_{i=1}^{k} \mathrm{I}(\Delta_{\mathrm{i}} \le \mathrm{s}) \tag{9}$$

and

$$\mathrm{I}(\Delta_i \le \mathrm{s}) = \begin{cases} 1 \dagger if \dagger \Delta_i \le s \\ 0 \dagger if \dagger \Delta_i > s \end{cases} \tag{10}$$

is the empirical distribution function.

Results

Comparison of three methods

Now the probabilities of the three methods for θ are compared. The difference between the sample proportions (horizontal axis) were plotted against the difference between the probabilities of the MCMC and exact methods (vertical axis), as shown in Figures 1, 3, and 5. Similarly, the difference between the sample proportions (horizontal axis) were plotted against the difference between the probabilities of the approximate and exact methods (vertical axis), as shown in Figures 2, 4, and 6. In Figures 1, and 2 consider small sample sizes, i.e., $\mathrm{n}_1 = \mathrm{n}_2 = 5$, 10, 15, and 20. Conversely, in Figures 3 and 4 consider large sample sizes, i.e., $\mathrm{n}_1 = \mathrm{n}_2 =$ 60, 70, 80, and 90. Figures 5 and 6 consider groups of different sample sizes, that is, $\mathrm{n}_1 = 15$, $\mathrm{n}_2 = 5$; $\mathrm{n}_1 = 15$, $\mathrm{n}_2 = 10$; $\mathrm{n}_1 = 15$, $\mathrm{n}_2 = 20$; and $\mathrm{n}_1 = 15$, $\mathrm{n}_2 = 20$. The following were confirmed from the results.

First, the relationship between the difference in the probabilities and the difference in the sample proportions is described. In Figure 1(d) and Figure 3(d), the probability of the MCMC method is more or less equal to that of the exact method when the difference between the sample proportions is 0.8. On the other hand, the difference between the probabilities of the MCMC and exact methods is around 0.01 when the difference between the sample proportions is 0.05. Overall, when the difference between the sample proportions is large, the probabilities of the MCMC and exact methods are roughly equal. In contrast, when the difference between the sample proportions is small, the probability of the MCMC method is

different from that of the exact method. This general pattern is similar for the difference in the probabilities of the approximation and exact methods.

Next, the relationship between the sample size and the difference in the probabilities is described. In Figure 2(a), the difference between the probabilities of the approximate and exact methods is around 0.013 when the difference between the sample proportions is 0.2. For a slightly larger sample size (Figure 2(d)), the difference between the probabilities of the approximate and exact methods is around 0.006 for the same difference between the sample proportions. In addition, there is virtually no difference between the probabilities of the approximate and exact methods when the sample size is further increased, as shown in Figure 4(d). Thus, the sample size influences the accuracy of the probability of the approximate method. It also shows the difference in the probabilities of the MCMC and exact methods. In Figure 1(a), the difference between the probabilities of the MCMC and exact methods is around 0.006 when the difference between the sample proportions is 0.2. For a slightly larger sample size (Figure 2(d)), the difference between the probabilities of the MCMC and exact methods is around 0.005 for the same difference between the sample proportions. Thus, the accuracy of the probability of the MCMC method always remains high even when the sample sizes are small.

Finally, the difference between the probabilities when groups of different sample sizes are considered is investigated. In Figure 2(d), the difference between the probabilities of the approximate and exact methods is around 0.006 when the difference between the sample proportions is 0.2. On the other hand, in Figure 6(d), the difference between the probabilities of the approximate and exact methods is around 0.012 for the same difference between the sample proportions. In both the cases, the total sample size ($n_1 + n_2$) is the same. However, the difference between the probabilities of the approximate and exact methods is slightly greater in the case of groups with different sample sizes. It is also shown the case of the MCMC method. In Figure 1(d), the difference between the probabilities of the MCMC and exact methods is around 0.005 when the difference between the sample proportions is 0.2. On the other hand, in Figure 5(d), the difference between the probability of the MCMC and exact methods is around 0.005 for the same difference between the sample proportions. Therefore, the difference between the probabilities of the MCMC and exact methods is the same regardless of whether the sample sizes are equal or different.

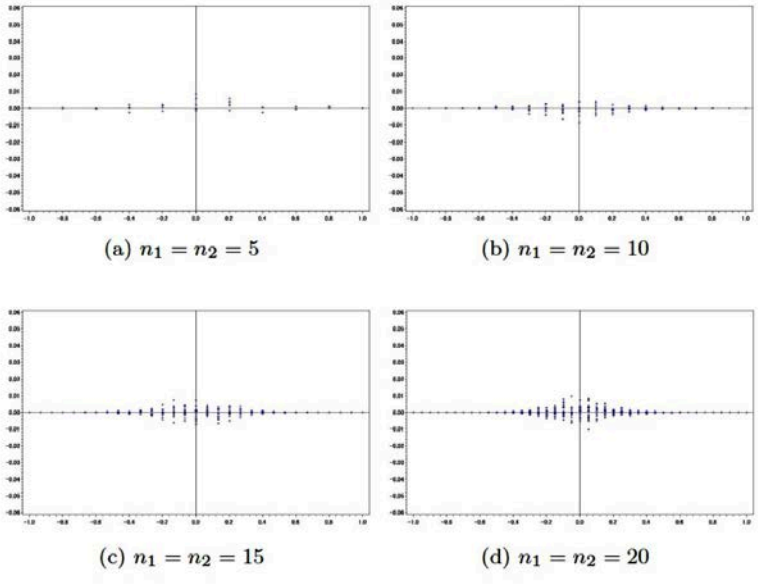


Figure 1: Comparison of the Exact and MCMC Method when sample sizes are small. (vertical axis：Differences of θ in Exact and MCMC method. Prior distribution is Beta(1,1). horizontal axis : Differences of two sample proportions.

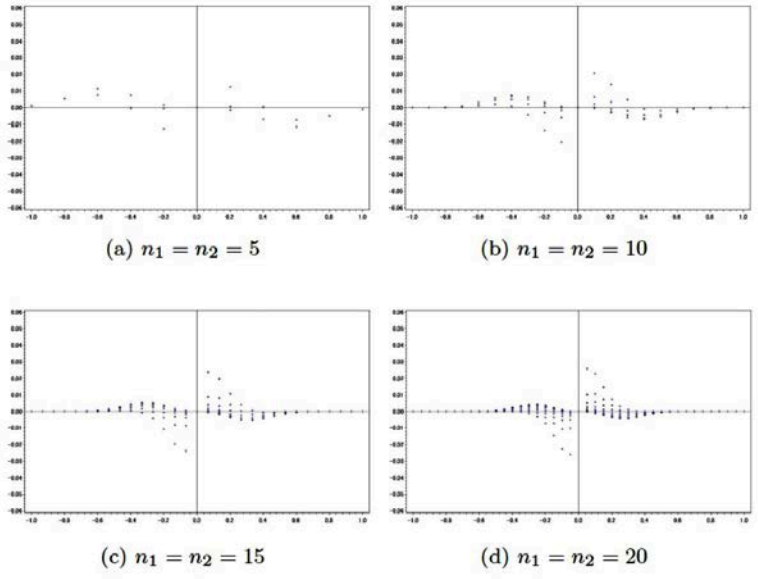


Figure 2: Comparison of the Exact and Approximate method when sample sizes are small. (vertical axis：Differences of θ in Exact and Approximation method. horizontal axis : Differences of two sample proportions.

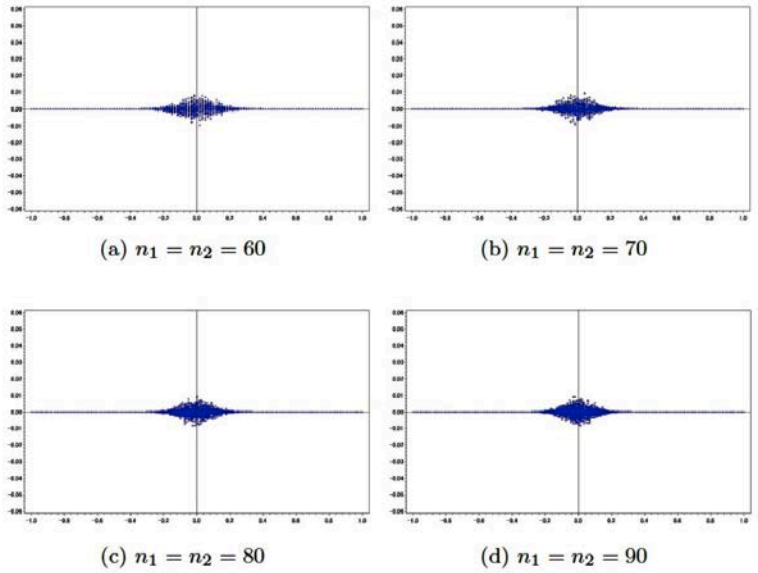


Figure 3: Comparison of the Exact and MCMC Method when sample sizes are large. (vertical axis：Differences of θ in Exact and MCMC method. Prior distribution is Beta(1,1). horizontal axis : Differences of two sample proportions.

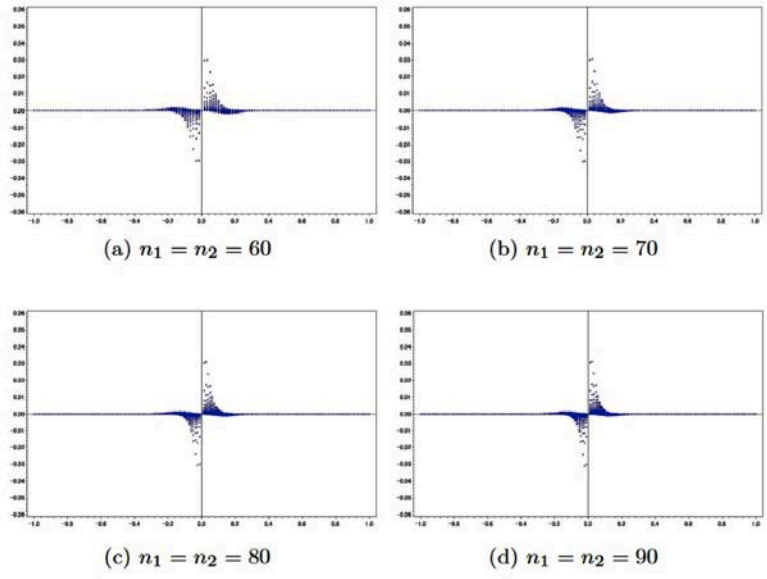


Figure 4: Comparison of the Exact and Approximate method when sample sizes are large. (vertical axis：Differences of θ in Exact and Approximation method. horizontal axis : Differences of two sample proportions.

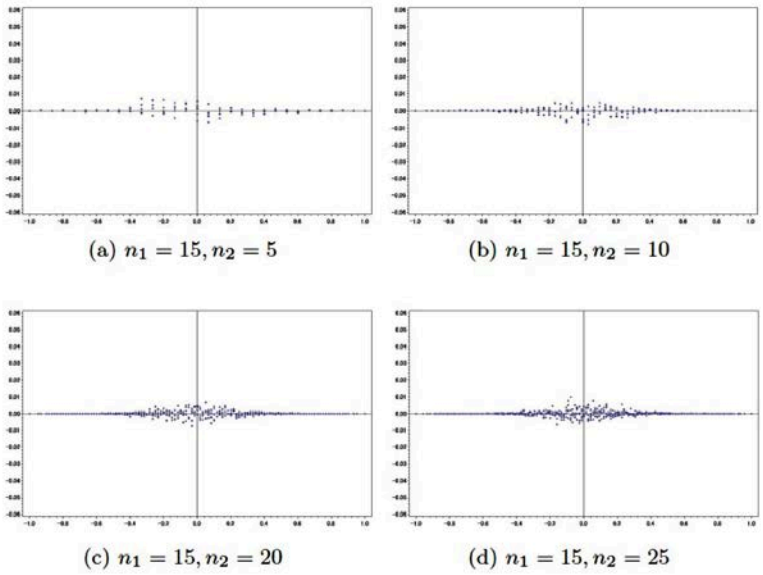


Figure 5: Comparison of the Exact and MCMC Method when sample sizes are unbalanced. (vertical axis : Differences of θ in Exact and MCMC method. Prior distribution is Beta(1,1). horizontal axis : Differences of two sample proportions.

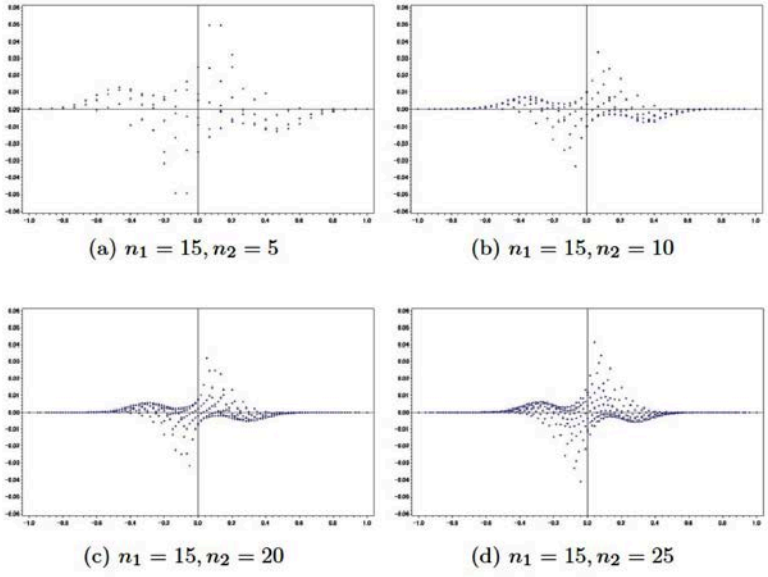


Figure 6: Comparison of the Exact and Approximate method when sample sizes are unbalanced. (vertical axis : Differences of θ in Exact and Approximation method. horizontal axis : Differences of two sample proportions.

Example

Next the utility of θ is illustrated by applying it to the results of clinical trials. A non-informative prior was assumed. Table 1 lists the results of a double-blind, randomized, 41-center study that compares the efficacy of TJN-318 cream with that of Bifonazole (BFZ) cream in the treatment of patients suffering from cutaneous mycosis (TJN-318 Solution Study Group (1992)). The main purpose of this clinical trial was to show that TJN-318 cream is more effective than BFZ cream in the treatment of cutaneous mycosis. The primary end point of this clinical trial is a binary variable. In other words, the patient either recovers or does not recover. In short, the alternative hypothesis is $\pi_1 > \pi_2$. In general, the frequentist approach can be adopted to verify the purpose of the clinical trial via the calculation of a *p*-value. The *p*-value was calculated using the Z-test statistic for the purpose of reference,

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}\left(1-\hat{\pi}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{11}$$

where $\hat{\pi}_i = x_i / n_i$ and $\hat{\pi} = \left(x_1 + x_2\right) / \left(n_1 + n_2\right)$. The values of θ are listed in the rightmost column of Table 1. Consequently, a non-informative prior was adopted, that is, $\alpha_i = \beta_i = 1$ and $i = 1, \dagger 2$. Clearly, θ increases when the *p*-value is low, and $\theta \approx 1$ when the null hypothesis is rejected. Moreover, $\theta \approx 1 - \dagger p$-value.

Next, the results of a double-blind, randomized, phase-3 clinical trial that compares the efficacies of follitropin alpha (hereafter, the study drug) and human menopausal gonadotropin (hereafter, the control drug) in the treatment of patients suffering from no-ovulation-cycle syndrome (from the assessment report of PMDA (2009)) was employed. Table 2 lists the resulting ovulation rate, that is, the primary end point. The Z-test affords a *p*-value of 0.764, which suggests no significant differences. Using the non-informative prior, the approximate probability of θ is obtained as 0.238, whereas the exact probability and the MCMC probability is obtained as 0.237.

Table 1: The result of primary end point in clinical trial for TJN-318 cream vs Bifonazole cream.

		Outcome			θ		
Disease Name	**Drug Name**	**Cure**	**Non-Cure**	***p*-value**	**Approximate**	**Exact**	**MCMC**
Tinea Pedis	TJN-318	110	27	0.264	0.734	0.735	0.735
	BFZ	96	31				
Tine Corporis	TJN-318	70	13	0.417	0.581	0.582	0.581
	BFZ	69	14				
Candidal Intertigo	TJN-318	39	4	0.472	0.531	0.530	0.531
	BFZ	37	4				
Candidal Interdigital	TJN-318	25	2	0.021	0.978	0.977	0.977
	BFZ	23	9				
Ptyriasis Versicolor	TJN-318	59	2	0.236	0.749	0.756	0.757
	BFZ	46	3				

Table 2: The result of primary end point in clinical trial for follitropin alpha vs human menopausal gonadotropin.

	Outcome					θ		
Drug Name	**Cure**	**Non-cure**	**Total**	**Ovulation Ratio**	***p*-value**	**Approximate**	**Exact**	**MCMC**
Study	102	27	129	79.1%	0.764	0.238	0.237	0.237
Control	109	23	132	82.6%				

Conclusion

Three methods for the index $\theta = \mathrm{P}(\pi_1 > \pi_2 \mid X_1, \dagger X_2)$ were presented to determine the probability that the binomial proportion for a study drug is superior to that for a control drug. In particular, a new procedure was described based on the MCMC method. The probabilities of these three methods were compared to test the relative effectiveness of each.

The expression for the exact method was presented, which includes a hypergeometric series. It is speculated that this series causes the decrease in calculation efficiency when the sample size is very large. In addition, hypergeometric series are not built into SAS, which is a statistical software

program frequently used in pharmaceutical development. Therefore, if SAS is used, a calculation program for hypergeometric series must be developed.

It is easy to calculate the probability for using the approximation method. This is an advantage when the approximate probability is used. Conversely, when the difference in the sample proportions is small and the sample sizes are unbalanced, the accuracy the approximation method is poor. That is, the accuracy of the probability of the approximation method depends on the sample size.

This study showed that the accuracy of the MCMC method was greater than that of the approximation method. Moreover, the probability of the MCMC method can be easily calculated using SAS. In addition, it is possible to use the non-conjugate prior for the prior distribution in the MCMC method. The authors consider this as one of the advantages of the MCMC method

References

Basu, S. (1996). Bayesian hypotheses testing using posterior density ratios. *Statistics and Probability Letters, 30,* 79-86.

Berry, D. S. (1995). *Statistics: a Bayesian perspective*. Duxbury Press, New York.

Kawasaki, Y. & Miyaoka, E. (2010). On confidence intervals for $P(X < Y)$. *Journal of the Japanese Society of Computational Statistics, 23,* 1-12.

Kawasaki, Y. & Miyaoka, E. (2012). A Bayesian inference of $P(\pi_1 > \pi_2)$ for two proportions. *Journal of Biopharmaceutical Statistics, 22,* 425-437.

Kotz, S., Lumelskii, Y. & Pensky, M. (2003). *The stress-strength model and its generalizations.* World Scientific Publishing, New Jersey.

Pharmaceuticals & Medical Devices Agency. (2009). *Follitropin alpha of assessment report.* PMDA, Tokyo, (in Japanese).

Sen, P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin, 10,* 1-18.

Sen, P. K. (1967). A note on asymptotically distribution-free confidence bounds for $P(X < Y)$, based on two independent samples. *Sankhya Series, A, 29,* 95-102.

Shirahata, S. (1993). Estimate of variance of Wilcoxon-Mann-Whitney statistic. *Journal of the Japanese Society of Computational Statistics, 6,* 1-10.

TJN-318 Solution Study Group. (1992). Double-blind study on TJN-318 cream and Bifonazole cream to the patient with the cutaneous mycosis. *The Nishinihon Journal of Dermatology, 54*, 977—992, (in Japanese).

Zaslavsky, B. G. (2009). Bayes models of clinical trials with dichotomous outcomes and sample size determination. *Statistics in Biopharmaceutical Research, 1,* 149-158.

Zaslavsky, B. G. (2010). Bayesian versus frequentist hypotheses testing in clinical trials with dichotomous and countable outcomes. *Journal of Biopharmaceutical Statistics, 20,* 985-997.

Zhou, W. (2008). Statistical inference for $\mathrm{P}(X < Y)$. *Statistics in Medicine, 27*, 257-279.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 269-292.

ISSN 1538 − 9472

On Bayesian Estimation and Predictions for Two-Component Mixture of the Gompertz Distribution

Navid Feroze
Allama Iqbal Open University
Islamabad, Pakistan

Muhammad Aslam
Quaid-i-Azam University
Islamabad, Pakistan

Mixtures models have received sizeable attention from analysts in the recent years. Some work on Bayesian estimation of the parameters of mixture models have appeared. However, the were restricted to the Bayes point estimation The methodology for the Bayesian interval estimation of the parameters for said models is still to be explored. This paper proposes the posterior interval estimation (along with point estimation) for the parameters of a two-component mixture of the Gompertz distribution. The posterior predictive intervals are also derived and evaluated. Different informative and non-informative priors are assumed under a couple of loss functions for the posterior analysis. A simulation study was carried out in order to make comparisons among different point and interval estimators. The applicability of the results is illustrated via a real life example.

Keywords: Bayes estimators, loss functions, posterior distributions, censoring, mixture densities

Introduction

The Gompertz distribution is used to model survival times, human mortality and actuarial tables. It has many real life applications, especially in medical and actuarial studies. The Gompertz distribution is also used as a survival model in reliability. It has an increasing hazard rate for the life of the systems. Due to its complicated form it has not received enough attention in past. However, recently, this distribution has received considerable attention from demographers and actuaries. Pollard and Valkovics (1992) were the first to deal with the Gompertz distribution thoroughly. However, their results are true only in cases where the initial level of mortality is very close to zero. Willemse and Koppelaar (2000)

Navid Feroze is a lecturer at the Government Post Graduate College Muzaffarābād, Azad Kashmir, Pakistan. Email him at: navidferoz@hotmail.com. Dr. Aslam is a professor in the Department of Statistics. Email him at aslamsdqu@yahoo.com.

reformulated the Gompertz force of mortality and derived relationships for this new formulation. Jaheen (2003) applied the Bayesian approach on record values from the Gompertz distribution. The simulation study was used for illustration of the results. Wu et al. (2003) derived the point and interval estimators for parameters of the Gompertz distribution under progressive type II censored samples. Wu et al. (2004) used the least square method to estimate the parameters of the Gompertz distribution. Wu et al. (2006) obtained the maximum likelihood estimators and the estimated expected test time for the two-parameter Gompertz distribution under progressive censoring with binomial removals. Khedhair and Gohary (2008) proposed the bivariate Gompertz distribution and completed the analysis for the mixture of components of proposed distribution. Saracoglu et al. (2009) compared the maximum likelihood, uniformly minimum variance unbiased, and Bayes estimators for the parameter of the Gompertz distribution. The numerical example was used for illustration. Ismail (2010) considered the Gompertz distribution as a lifetime model for applying the Bayesian approach to the estimation problem in the case of step stress partially accelerated life tests with two stress levels and type-I censoring. Ismail (2011) discussed the point and interval estimations of a two-parameter Gompertz distribution under partially accelerated life tests with Type-II censoring. Kiani et al. (2012) studied the performance of the Gompertz model with time-dependent covariate in the presence of right censored data. Moreover, the performance of the model was compared at different censoring proportions (*CP*) and sample sizes.

The mixture models have received great interest from analysts in recent era. These models include finite and infinite numbers of components that can analyze different datasets. A finite mixture of probability distribution is suitable to study a population categorized in number of subpopulations. A population of lifetimes of certain electrical elements can be classified into a number of subpopulations based on causes of failures. The analysis of mixture models under Bayesian framework has developed a significant interest among statisticians. Authors dealing with the Bayesian analysis of mixture models include: Saleem and Aslam (2008), Saleem et al. (2010), Majeed and Aslam (2012) and Kazmi et al. (2012). These contributions are concerned with Bayes point estimation of the parameters. The interval estimation of the parameters of the mixture models under a Bayesian framework has not yet been discussed by any author. We considered point and interval estimation of the parameters for a two-component mixture of the Gompertz distribution. The population of certain items is assumed to be partitioned into two subpopulations. The randomly selected observations from said population are considered to be a part of one of the above mentioned

subpopulations. These subpopulations are assumed to follow the Gompertz distribution. Therefore, the two components mixture of the Gompertz distribution has been proposed to model this population. The observations have been assumed to be right censored. The inverse transformation technique of simulation under a probabilistic mixing has been used to generate data and to evaluate the performance of different estimators.

The Population and the Model

A density function for the mixture of two component densities with mixing weights (p,q) is

$$f(x) = pf_1(x) + qf_2(x), 0 < p < 1 \tag{1}$$

The following Gompertz distribution is considered for both mixture densities:

$$f(x_i;\alpha_i) = \alpha_i e^{x_i - \alpha_i\left(e^{x_i}-1\right)}, x_i > 0, \alpha_i > 0 \tag{2}$$

with the cumulative distribution function as

$$F(x_i;\alpha_i) = 1 - e^{-\alpha_i\left(e^{x_i}-1\right)} \tag{3}$$

The cumulative distribution function for the mixture model is

$$F(x) = pF_1(x) + qF_2(x) \tag{4}$$

Suppose n items are put on a life testing experiment and w units failed until time T, while $n - w$ units are still working. Now based on causes of failure, the failed items are assumed to come either from subpopulation 1 or from subpopulation 2. Therefore it can be observed that w_1 and w_2 failed items come from the first and second subpopulation respectively, where $w = w_1 + w_2$. The remaining $n - w$ items are assumed to be censored observations. The likelihood function for above type I censored data can be obtained as

$$L\left(\alpha_1,\alpha_2,p\left|\underline{x}\right.\right)\propto\prod_{j=1}^{w_1}\left\{pf_1\left(x_{1j}\right)\right\}\prod_{j=1}^{w_2}\left\{pf_2\left(x_{2j}\right)\right\}\times\left[1-F\left(t\right)\right]^{n-w} \tag{5}$$

After simplifications the likelihood function becomes

$$L\left(\alpha_1,\alpha_2,p\left|\underline{x}\right.\right)\propto\sum_{k=0}^{n-w}\binom{n-w}{k}\alpha_1^{w_1}\alpha_2^{w_2}p^{n-k-w_2}\times q^{w_2+k}e^{-\alpha_1\xi_{1k}(x)}e^{-\alpha_2\xi_{2k}(x)} \tag{6}$$

where

$$\xi_{1k}\left(x\right)=\sum_{j=1}^{w_1}\left(e^{x_{1j}}-1\right)+\left(n-w-k\right)\left(e^t-1\right)$$

and

$$\xi_{2k}\left(x\right)=\sum_{j=1}^{w_2}\left(e^{x_{2j}}-1\right)+k\left(e^t-1\right)$$

The Posterior Distributions under Different Priors

The main difference between the Bayesian and classical inference is the use of prior information under the Bayesian framework. However, in cases where the sufficient prior information regarding the parameter is not available, the use of non-informative priors becomes mandatory. An important non-informative prior, proposed by Laplace (1812), is a uniform prior. It has been applied to many problems, and often the results are entirely satisfactory. Here, this prior has been used for the posterior estimation.

Let $\alpha_1\in\text{Uniform}\ \forall\ \alpha_1\in\left(0,\infty\right)$, $\alpha_2\in\text{Uniform}\ \forall\ \alpha_2\in\left(0,\infty\right)$, and $p\sim U\left(0,1\right)$. Assuming independence, these priors result into a joint prior that is proportional to a constant. That joint prior has been used to derive the joint posterior distribution of α_1,α_2 and p. The marginal distribution for each

parameter can be obtained by integrating the joint posterior distribution with respect to nuisance parameters. The joint posterior distribution is

$$p\left(\alpha_1,\alpha_2,p\middle|\underline{x}\right)=\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\alpha_1^{w_1}\alpha_2^{w_2}p^{n-k-w_2}q^{w_2+k}e^{-\alpha_1\xi_{1k}(x)}e^{-\alpha_2\xi_{2k}(x)}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1}}\frac{\Gamma\left(w_2+1\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1}}},\qquad(7)$$

$$a_1,a_2>0$$

where $\psi_{1k}=n-k-w_2+1$, $\psi_{2k}=w_2+k+1$ and $B\left(\psi_{1k},\psi_{2k}\right)$ is standard beta function.

Another non-informative prior has been suggested by Jeffreys (1961), and is frequently used in situations where one does not have much information about the parameters. This prior is defined as

$$p\left(\underline{\alpha}\right)\propto\left\{\left|I\left(\underline{\alpha}\right)\right|\right\}^{\frac{1}{2}}$$

where $f_i\left(x_i\middle|\alpha_i\right)$ have been defined in (2) and $p\sim U\left(0,1\right)$. Assuming independence, the joint prior is obtained as

$$h\left(\alpha_1,\alpha_2,p\right)\propto\frac{1}{\alpha_1\alpha_2},a_1,a_2>0\qquad(8)$$

The joint posterior distribution using the above prior is

$$p\left(\alpha_1,\alpha_2,p\left|\underline{x}\right.\right)=\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\alpha_1^{w_1-1}\alpha_2^{w_2-1}p^{n-k-w_2}q^{w_2+k}e^{-\alpha_1\xi_{1k}(x)}e^{-\alpha_2\xi_{2k}(x)}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1}}\frac{\Gamma\left(w_2\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2}}},\quad a_1,a_2>0 \tag{9}$$

The utilization of informative prior is of much importance under Bayesian inference. The results under informative priors are often better than non-informative priors. The gamma, chi square and exponential priors have been assumed for the posterior analysis in the current study. The combined priors have been obtained by assuming the independence.

Let $\alpha_1:\text{Gamma}\left(\sigma_1,\tau_1\right)$, $\alpha_2\sim$ Gamma$\left(\sigma_2,\tau_2\right)$ and $p\sim$ Uniform$\left(0,1\right)$. Under the assumption of independence, the joint prior becomes

$$h\left(\alpha_1,\alpha_2,p\right)\propto\alpha_1^{\sigma_1-1}\alpha_2^{\sigma_2-1}e^{-\left(\alpha_1\tau_1+\alpha_2\tau_2\right)},a_1,a_2>0 \tag{10}$$

The posterior distribution under the assumption of above prior is

$$p\left(\alpha_1,\alpha_2,p\left|\underline{x}\right.\right)=\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\alpha_1^{w_1+\sigma_1-1}\alpha_2^{w_2+\sigma_2-1}p^{n-k-w_2}q^{w_2+k}e^{-\alpha_1\left\{\xi_{1k}(x)+\tau_1\right\}}e^{-\alpha_2\left\{\xi_{2k}(x)+\tau_2\right\}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+\sigma_1\right)}{\left\{\xi_{1k}\left(x\right)+\tau_1\right\}^{w_1+\sigma_1}}\frac{\Gamma\left(w_2+\sigma_2\right)}{\left\{\xi_{2k}\left(x\right)+\tau_2\right\}^{w_2+\sigma_2}}},\quad a_1,a_2>0 \tag{11}$$

Again, suppose $\alpha_1\sim$ Chi Square$\left(\nu_1\right)$, $\alpha_2\sim$ Chi Square$\left(\nu_2\right)$ and $p\sim$ Uniform$\left(0,1\right)$. Assuming independence, the joint prior becomes

$$h\left(\alpha_1,\alpha_2,p\right)\propto\alpha_1^{\frac{\nu_1}{2}-1}\alpha_2^{\frac{\nu_2}{2}-1}e^{-\frac{\left(\alpha_1+\alpha_2\right)}{2}},a_1,a_2>0 \tag{12}$$

The posterior distribution under the assumption of the prior given in (12) is

$$p\left(\alpha_1,\alpha_2,p\left|\underline{x}\right.\right)=\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\alpha_1^{w_1+0.5\nu_1-1}\alpha_2^{w_2+0.5\nu_2-1}p^{n-k-w_2}q^{w_2+k}e^{-\alpha_1\left\{\xi_{1k}(x)+0.5\right\}}e^{-\alpha_2\left\{\xi_{2k}(x)+0.5\right\}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+0.5\nu_1\right)}{\left\{\xi_{1k}\left(x\right)+0.5\right\}^{w_1+0.5\nu_1}}\frac{\Gamma\left(w_2+0.5\nu_2\right)}{\left\{\xi_{2k}\left(x\right)+0.5\right\}^{w_2+0.5\nu_2}}},\quad(13)$$

$$a_1,a_2>0$$

Further, consider $\alpha_1 \sim$ Exponential$\left(\varphi_1\right)$, $\alpha_2 \sim$ Exponential$\left(\varphi_2\right)$ and $p \sim$ Uniform$\left(0,1\right)$. Under the assumption of independence, the joint prior becomes

$$h\left(\alpha_1,\alpha_2,p\right)\propto e^{-\left(\alpha_1\varphi_1+\alpha_2\varphi_2\right)},a_1,a_2>0 \quad(14)$$

The posterior distribution under the assumption of above prior is

$$p\left(\alpha_1,\alpha_2,p\left|\underline{x}\right.\right)=\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\alpha_1^{w_1}\alpha_2^{w_2}p^{n-k-w_2}q^{w_2+k}e^{-\alpha_1\left\{\xi_{1k}(x)+\varphi_1\right\}}e^{-\alpha_2\left\{\xi_{2k}(x)+\varphi_2\right\}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)}{\left\{\xi_{1k}\left(x\right)+\varphi_1\right\}^{w_1+1}}\frac{\Gamma\left(w_2+1\right)}{\left\{\xi_{2k}\left(x\right)+\varphi_2\right\}^{w_2+1}}},\quad(15)$$

$$a_1,a_2>0$$

Bayes Estimators and Posterior Risks

The Bayes estimators and associated posterior risks have been derived under the squared error loss function (*SELF*) and precautionary loss function (*PLF*). The respective expressions have been presented in the following.

Bayes estimator and posterior risk for α_1, α_2 and p under uniform prior using *SELF* are:

$$\left(B.E\right)_{PLF}=\left[\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k}+2h,\psi_{2k}\right)\frac{\Gamma\left(w_1+1+2i\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1+2i}}\frac{\Gamma\left(w_2+1+2j\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1+2j}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1}}\frac{\Gamma\left(w_2+1\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1}}}\right]^{\frac{1}{2}}$$

$$\rho\left\{\left(B.E\right)_{PLF}\right\}=2\left[\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k}+2h,\psi_{2k}\right)\frac{\Gamma\left(w_1+1+2i\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1+2i}}\frac{\Gamma\left(w_2+1+2j\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1+2j}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1}}\frac{\Gamma\left(w_2+1\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1}}}\right]^{\frac{1}{2}}$$
$$-\frac{2\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k}+h,\psi_{2k}\right)\frac{\Gamma\left(w_1+1+i\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1+i}}\frac{\Gamma\left(w_2+1+j\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1+j}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)}{\left\{\xi_{1k}\left(x\right)\right\}^{w_1+1}}\frac{\Gamma\left(w_2+1\right)}{\left\{\xi_{2k}\left(x\right)\right\}^{w_2+1}}}$$

where, $\left(B.E\right)_{PLF}$ and $\rho\left\{\left(B.E\right)_{PLF}\right\}$ are the Bayes estimator and the posterior risk under *PLF*. The Bayes estimates and corresponding risks under other priors can be derived in the similar manner.

Credible intervals

The credible interval is defined as: Let $g\left(\alpha|\mathrm{x}\right)$ be the posterior distribution then a $100\left(1-k\right)\%$ credible interval in any set C is such that $P_{g\left(\alpha|\mathrm{x}\right)}\left(C\right)=1-k$. According to Eberly and Casella (2003) the credible interval can also be defined as: $\int_0^L g\left(\alpha|\mathrm{x}\right)d\alpha=\frac{k}{2}, \int_U^\infty g\left(\alpha|\mathrm{x}\right)d\alpha=\frac{k}{2}$ where L and U are the lower and upper limits of the credible interval respectively and *k* is level of significance.

The $100(1-k)\%$ credible intervals for α_1, α_2 and p under uniform prior can be obtained by solving the following two equations.

$$\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(L^h,\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1,iL\xi_{1k}(x)\right)\Gamma\left(w_2+1,jL\xi_{2k}(x)\right)}{\left\{\xi_{1k}(x)\right\}^{w_1+1}\left\{\xi_{2k}(x)\right\}^{w_2+1}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)\Gamma\left(w_2+1\right)}{\left\{\xi_{1k}(x)\right\}^{w_1+1}\left\{\xi_{2k}(x)\right\}^{w_2+1}}}=1-\frac{k}{2}$$

$$\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(U^h,\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1,iU\xi_{1k}(x)\right)\Gamma\left(w_2+1,jU\xi_{2k}(x)\right)}{\left\{\xi_{1k}(x)\right\}^{w_1+1}\left\{\xi_{2k}(x)\right\}^{w_2+1}}}{\sum_{k=0}^{n-w}\binom{n-w}{k}B\left(\psi_{1k},\psi_{2k}\right)\frac{\Gamma\left(w_1+1\right)\Gamma\left(w_2+1\right)}{\left\{\xi_{1k}(x)\right\}^{w_1+1}\left\{\xi_{2k}(x)\right\}^{w_2+1}}}=\frac{k}{2}$$

where h, i, j = 0.1, $\Gamma(x,y)$ is incomplete gamma function, $B(x,y,z)$ is incomplete beta function and (L,U) define the limits of the credible intervals. Now, the credible interval for α_1, α_2 and p can be derived by putting $h=0, i=1, j=0$, $h=0, i=0, j=1$, and $h=1, i=0, j=0$, respectively, in the above equations. It should be noted that $\Gamma(x,0)=\Gamma(x)$ and $B(1,y,z)=B(y,z)$. It can be observed that the explicit solution of the limits for the credible intervals cannot be obtained. The numerical methods have been used to find the approximate solution of the limits.

Posterior Predictive Distributions and Intervals

The posterior predictive distribution is used to make predictions of future observations, based on the inferences drawn from the data at hand. Posterior predictive distribution can be simply obtained by the product of the posterior distribution and (conditional) independence (given the parameters) of the new observation from the current sample. It can be defined as

$$g(y|\mathrm{x})=\int_0^\infty\int_0^\infty\int_0^1 p(\alpha_1,\alpha_2,p|\mathrm{x}) \times f(y;\alpha_1,\alpha_2,p)\,dp\,d\alpha_1\,d\alpha_2 \tag{16}$$

where y = x_{n+1} is the future observation given the sample information $x = x_1, x_2, ..., x_n$, from the model (7). The posterior predictive distribution using (7) and (16) can be obtained as

$$g(y|\mathrm{x}) = \frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\left[\frac{B(\psi_{1k}+1,\psi_{2k})(w_1+1)e^{y}}{\{\xi_{1k}(x)-\ln(e^{y}-1)\}^{w_1+2}\{\xi_{2k}(x)\}^{w_2+1}}+\frac{B(\psi_{1k},\psi_{2k}+1)(w_2+1)e^{y}}{\{\xi_{1k}(x)\}^{w_1+1}\{\xi_{2k}(x)-\ln(e^{y}-1)\}^{w_2+2}}\right]}{\sum_{k=0}^{n-w}\binom{n-w}{k}\frac{B(\psi_{1k},\psi_{2k})}{\{\xi_{1k}(x)\}^{w_1+1}\{\xi_{2k}(x)\}^{w_2+1}}} \quad (17)$$

The posterior predictive interval can be obtained by solving the following two equations

$$\int_0^L g(y|\mathrm{x})dy = \frac{\alpha}{2},\ \int_U^{\infty} g(y|\mathrm{x})dy = \frac{\alpha}{2}$$

The simplification of the above equations leads to the following equations

$$\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\left[\frac{B(\psi_{1k}+1,\psi_{2k})}{\{\xi_{2k}(x)\}^{w_2+1}}\left[\{\xi_{1k}(x)\}^{-w_1-1}-\{\xi_{1k}(x)-(e^{L}-1)\}^{-w_1-1}\right]+\frac{B(\psi_{1k},\psi_{2k}+1)}{\{\xi_{1k}(x)\}^{w_1+1}}\left[\{\xi_{2k}(x)\}^{-w_2-1}-\{\xi_{2k}(x)-(e^{L}-1)\}^{-w_2-1}\right]\right]}{\sum_{k=0}^{n-w}\binom{n-w}{k}\frac{B(\psi_{1k},\psi_{2k})}{\{\xi_{1k}(x)\}^{w_1+1}\{\xi_{2k}(x)\}^{w_2+1}}} = \frac{k}{2}$$

$$\frac{\sum_{k=0}^{n-w}\binom{n-w}{k}\left[\frac{B(\psi_{1k}+1,\psi_{2k})}{\{\xi_{2k}(x)\}^{w_2+1}}\left[\{\xi_{1k}(x)\}^{-w_1-1}-\{\xi_{1k}(x)-(e^{U}-1)\}^{-w_1-1}\right]+\frac{B(\psi_{1k},\psi_{2k}+1)}{\{\xi_{1k}(x)\}^{w_1+1}}\left[\{\xi_{2k}(x)\}^{-w_2-1}-\{\xi_{2k}(x)-(e^{U}-1)\}^{-w_2-1}\right]\right]}{\sum_{k=0}^{n-w}\binom{n-w}{k}\frac{B(\psi_{1k},\psi_{2k})}{\{\xi_{1k}(x)\}^{w_1+1}\{\xi_{2k}(x)\}^{w_2+1}}} = 1-\frac{k}{2}$$

As the limits of the posterior predictive interval cannot be derived explicitly, the numerical solutions of the limits have been obtained by iterative methods.

Prior Elicitation

The elicitation is a technique to formulate an expert's knowledge or belief about a certain quantity into a joint probability distribution. In the case of Bayesian analysis, it can be considered as a method to specify the values of hyper-parameters in a prior distribution for one or more parameters of the sampling distribution.

Much of the literature on elicitation has been concerned with formulating a probability distribution for unsure quantities when there is no data with which to supplement the knowledge expressed in that distribution. This process facilitates decision-making, where uncertainty about certain phenomena needs to be described in terms of a probability distribution in order to derive the posterior distributions.

To achieve accurate elicitation is a difficult task, even if we are interested in elicitation of a single event. In such a situation, a single probability is needed, but the expert may not be familiar with the concept of probabilities. Even when the expert is familiar with the concept of probabilities, it is by no means straightforward to evaluate a probability value for an event exactly. In such cases, elicitation encourages the expert and the facilitator to consider the meaning of the parameters being elicited. This has two helpful consequences. First, it brings the analysis closer to the application by demanding attention to what is being modeled, and what is reasonable to believe about it. Second, it helps to make the posterior distributions, once calculated, into meaningful quantities. Many methods of elicitation have been discussed in the literature; among those, the method suggested by Aslam (2003) has been used to elicit the prior distribution in the recent study. This method requires the derivation of prior predictive distribution for elicitation. The prior predictive distribution can be defined as

$$g(y)=\int_0^{\infty}\int_0^{\infty}\int_0^{1} h(\alpha_1,\alpha_2,p)\, f\left(y\middle|\alpha_1,\alpha_2,p\right) dp\, d\alpha_1 d\alpha_2 \tag{18}$$

where $h(\alpha_1,\alpha_2,p)$ and $f\left(y\middle|\alpha_1,\alpha_2,p\right)$ are prior distribution and mixture Gompertz model respectively.

According to (18), the prior predictive distribution under gamma prior is

$$g(y)=\frac{\tau_1^{\sigma_1}\tau_2^{\sigma_2}e^y}{2\Gamma(\sigma_1)\Gamma(\sigma_2)}\left[\frac{\Gamma(\sigma_1+1)\Gamma(\sigma_2)}{(e^y+\tau_1-1)^{\sigma_1+1}(e^y-1)^{\sigma_2}}+\frac{\Gamma(\sigma_1)\Gamma(\sigma_2+1)}{(e^y-1)^{\sigma_1}(e^y+\tau_2-1)^{\sigma_2+1}}\right]$$

In order to elicit the four hyper-parameters, the following four integrals have been considered. The expert's probabilities have been assumed to be 0.15 for each integral:

$$\int_1^{20} g(y)=0.15,\ \int_{21}^{40} g(y)=0.15,\ \int_{41}^{60} g(y)=0.15 \text{ and } \int_{61}^{80} g(y)=0.15.$$

A program has been developed in SAS package using the "PROC SYSLIN" command to solve the above integrals simultaneously. The set of hyper-parameters with minimum values has been chosen to be the elicited values of the hyper-parameters. These elicited values of the hyper-parameters have been found to be $(\sigma_1,\tau_1,\sigma_2,\tau_2)=(0.000233,0.190642,0.000101,0.189112)$. The prior predictive distribution under chi square prior, given in (12), has been derived as

$$g(y)=\frac{e^y}{2^{0.5(\nu_1+\nu_2)+1}\Gamma(0.5\nu_1)\Gamma(0.5\nu_2)}\left[\frac{\Gamma(0.5\nu_1+1)\Gamma(0.5\nu_2)}{(e^y-0.5)^{0.5\nu_1+1}(e^y-1)^{0.5\nu_2}}+\frac{\Gamma(0.5\nu_1)\Gamma(0.5\nu_2+1)}{(e^y-1)^{0.5\nu_1}(e^y-0.5)^{0.5\nu_2+1}}\right] \quad (19)$$

Using the similar program mentioned above, the elicited values of the hyper-parameters are $(\nu_1,\nu_2)=(1.226759,1.064564)$.

The prior predictive distribution under chi square prior, given in (14), has been presented as

$$g(y)=0.5\varphi_1\varphi_2e^y(e^y-1)^{-1}\left[(e^y+\varphi_1-1)^{-2}+(e^y+\varphi_2-1)^{-2}\right] \quad (20)$$

The elicited values of the hyper-parameters in the above prior predictive distribution are $(\varphi_1,\varphi_2)=(0.232768,0.322483)$.

Results and Discussion

A simulation study has been conducted to assess and compare the performance of Bayes estimators and to analyze the impact of sample size, mixing weight and magnitude of parametric values on the Bayes estimators. Samples of sizes n = 50, 100, 200, 300, 400 and 500 have been generated by inverse transformation method from a two components mixture of the Gompertz distribution. The parametric values used are: $(\alpha_1, \alpha_2) \in \{(4,6),(8,12)\}$ and $p \in (0.45, 0.60)$. Probabilistic mixing has been used to generate the mixture data. For each observation a random number u has been generated from $U(0,1)$. If $u < p$ the observation has been randomly taken from first subpopulation and if $u > p$ then the observation has been taken from the second subpopulation. The observations above a fixed censoring time T have been assumed to be right censored. Under each combination of parametric values, the choice of censoring time has been made so that the censoring rate in the respective sample is 15%. As one sample cannot completely describe the behavior and properties of the Bayes estimators, the results have been replicated 1000 times and the average of results has been presented in the tables below (the amounts of posterior risks are presented in parenthesis). The abbreviations used in tables are; *B.Es*: Bayes estimates; *P.Rs*: posterior risks; *LL*: lower limit and *UL*: upper limit.

Table 1. *B.Es* and *P.Rs* under Uniform Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 4, p = 0.45$		$\alpha_1 = 4, p = 0.60$		$\alpha_2 = 6, p = 0.45$		$\alpha_2 = 6, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	4.5131	4.5363	4.3991	4.4217	6.7921	6.8271	6.7307	6.7653	0.5190	0.5217	0.6709	0.6743
	(0.2039)	(0.0464)	(0.1987)	(0.0452)	(0.3068)	(0.0698)	(0.3040)	(0.0692)	(0.0612)	(0.0139)	(0.0845)	(0.0192)
100	4.4795	4.4908	4.3664	4.3775	6.7416	6.7587	6.6806	6.6975	0.5151	0.5164	0.6659	0.6676
	(0.0989)	(0.0227)	(0.0964)	(0.0221)	(0.1488)	(0.0342)	(0.1475)	(0.0339)	(0.0297)	(0.0068)	(0.0410)	(0.0094)
200	4.3936	4.3991	4.2827	4.2881	6.6124	6.6207	6.5525	6.5608	0.5053	0.5059	0.6531	0.6539
	(0.0472)	(0.0111)	(0.0460)	(0.0108)	(0.0710)	(0.0166)	(0.0704)	(0.0165)	(0.0142)	(0.0033)	(0.0196)	(0.0046)
300	4.2706	4.2742	4.1628	4.1663	6.4273	6.4327	6.3691	6.3744	0.4911	0.4915	0.6348	0.6354
	(0.0297)	(0.0072)	(0.0289)	(0.0070)	(0.0446)	(0.0108)	(0.0442)	(0.0107)	(0.0089)	(0.0021)	(0.0123)	(0.0030)
400	4.1707	4.1734	4.0655	4.0680	6.2770	6.2809	6.2202	6.2241	0.4796	0.4799	0.6200	0.6204
	(0.0212)	(0.0052)	(0.0207)	(0.0051)	(0.0319)	(0.0079)	(0.0316)	(0.0078)	(0.0064)	(0.0016)	(0.0088)	(0.0022)
500	4.1428	4.1449	4.0383	4.0403	6.2350	6.2381	6.1785	6.1816	0.4764	0.4767	0.6158	0.6161
	(0.0167)	(0.0042)	(0.0163)	(0.0040)	(0.0252)	(0.0063)	(0.0249)	(0.0062)	(0.0050)	(0.0012)	(0.0069)	(0.0017)

Table 2. *B.Es* and *P.Rs* under Uniform Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 8, p = 0.45$		$\alpha_1 = 8, p = 0.60$		$\alpha_2 = 12, p = 0.45$		$\alpha_2 = 12, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	8.8508	8.8963	8.8947	8.9404	13.3205	13.3890	13.4754	13.5447	0.5311	0.5338	0.7027	0.7063
	(0.3998)	(0.0910)	(0.3563)	(0.0811)	(0.6017)	(0.1370)	(0.6769)	(0.1541)	(0.0760)	(0.0173)	(0.0994)	(0.0226)
100	8.7850	8.8072	8.8284	8.8508	13.2214	13.2549	13.3751	13.4090	0.5271	0.5284	0.6974	0.6992
	(0.1939)	(0.0445)	(0.1728)	(0.0397)	(0.2919)	(0.0670)	(0.3283)	(0.0754)	(0.0368)	(0.0085)	(0.0482)	(0.0111)
200	8.6166	8.6274	8.6592	8.6701	12.9679	12.9842	13.1187	13.1352	0.5170	0.5176	0.6841	0.6849
	(0.0926)	(0.0217)	(0.0825)	(0.0193)	(0.1393)	(0.0326)	(0.1567)	(0.0367)	(0.0176)	(0.0041)	(0.0230)	(0.0054)
300	8.3754	8.3824	8.4168	8.4239	12.6050	12.6155	12.7515	12.7622	0.5025	0.5029	0.6649	0.6655
	(0.0582)	(0.0140)	(0.0518)	(0.0125)	(0.0875)	(0.0211)	(0.0985)	(0.0237)	(0.0111)	(0.0027)	(0.0145)	(0.0035)
400	8.1795	8.1847	8.2200	8.2252	12.3102	12.3179	12.4533	12.4611	0.4908	0.4911	0.6494	0.6498
	(0.0416)	(0.0103)	(0.0370)	(0.0091)	(0.0625)	(0.0154)	(0.0704)	(0.0174)	(0.0079)	(0.0019)	(0.0103)	(0.0026)
500	8.1248	8.1289	8.1650	8.1691	12.2278	12.2339	12.3700	12.3762	0.4875	0.4877	0.6450	0.6454
	(0.0328)	(0.0081)	(0.0292)	(0.0073)	(0.0493)	(0.0123)	(0.0555)	(0.0138)	(0.0062)	(0.0015)	(0.0081)	(0.0020)

Table 3. *B.Es* and *P.Rs* under Jeffreys Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 4, p = 0.45$		$\alpha_1 = 4, p = 0.60$		$\alpha_2 = 6, p = 0.45$		$\alpha_2 = 6, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	4.4679	4.4909	4.3551	4.3775	6.7242	6.7588	6.6634	6.6976	0.5138	0.5165	0.6642	0.6676
	(0.2018)	(0.0459)	(0.1967)	(0.0448)	(0.3037)	(0.0691)	(0.3010)	(0.0685)	(0.0550)	(0.0125)	(0.0760)	(0.0173)
100	4.4347	4.4459	4.3227	4.3337	6.6742	6.6911	6.6138	6.6305	0.5100	0.5113	0.6592	0.6609
	(0.0979)	(0.0225)	(0.0954)	(0.0219)	(0.1473)	(0.0338)	(0.1460)	(0.0335)	(0.0267)	(0.0061)	(0.0369)	(0.0085)
200	4.3497	4.3551	4.2399	4.2452	6.5462	6.5545	6.4870	6.4951	0.5002	0.5008	0.6466	0.6474
	(0.0467)	(0.0109)	(0.0456)	(0.0107)	(0.0703)	(0.0165)	(0.0697)	(0.0163)	(0.0127)	(0.0030)	(0.0176)	(0.0041)
300	4.2279	4.2314	4.1212	4.1246	6.3630	6.3683	6.3054	6.3107	0.4862	0.4866	0.6285	0.6290
	(0.0294)	(0.0071)	(0.0286)	(0.0069)	(0.0442)	(0.0107)	(0.0438)	(0.0106)	(0.0080)	(0.0019)	(0.0111)	(0.0027)
400	4.1290	4.1316	4.0248	4.0273	6.2142	6.2181	6.1580	6.1618	0.4748	0.4751	0.6138	0.6142
	(0.0210)	(0.0052)	(0.0204)	(0.0050)	(0.0316)	(0.0078)	(0.0313)	(0.0077)	(0.0057)	(0.0014)	(0.0079)	(0.0020)
500	4.1014	4.1035	3.9979	3.9999	6.1726	6.1757	6.1167	6.1198	0.4717	0.4719	0.6097	0.6100
	(0.0165)	(0.0041)	(0.0161)	(0.0040)	(0.0249)	(0.0062)	(0.0247)	(0.0061)	(0.0045)	(0.0011)	(0.0062)	(0.0015)

Table 4. *B.Es* and *P.Rs* under Jeffreys Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 8, p = 0.45$		$\alpha_1 = 8, p = 0.60$		$\alpha_2 = 12, p = 0.45$		$\alpha_2 = 12, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	8.7623	8.8074	8.8057	8.8510	13.1873	13.2551	13.3406	13.4092	0.5257	0.5284	0.6957	0.6992
	(0.3958)	(0.0901)	(0.3527)	(0.0803)	(0.5957)	(0.1356)	(0.6701)	(0.1525)	(0.0684)	(0.0156)	(0.0895)	(0.0204)
100	8.6971	8.7191	8.7402	8.7623	13.0891	13.1223	13.2413	13.2749	0.5218	0.5231	0.6905	0.6922
	(0.1920)	(0.0441)	(0.1711)	(0.0393)	(0.2890)	(0.0664)	(0.3251)	(0.0747)	(0.0332)	(0.0076)	(0.0434)	(0.0100)
200	8.5304	8.5411	8.5726	8.5834	12.8382	12.8544	12.9875	13.0039	0.5118	0.5125	0.6772	0.6781
	(0.0917)	(0.0215)	(0.0817)	(0.0191)	(0.1379)	(0.0323)	(0.1552)	(0.0364)	(0.0158)	(0.0037)	(0.0207)	(0.0049)
300	8.2916	8.2986	8.3327	8.3397	12.4789	12.4893	12.6240	12.6346	0.4975	0.4979	0.6583	0.6588
	(0.0576)	(0.0139)	(0.0513)	(0.0124)	(0.0867)	(0.0209)	(0.0975)	(0.0235)	(0.0099)	(0.0024)	(0.0130)	(0.0031)
400	8.0977	8.1028	8.1378	8.1429	12.1871	12.1947	12.3288	12.3365	0.4859	0.4862	0.6429	0.6433
	(0.0411)	(0.0102)	(0.0367)	(0.0091)	(0.0619)	(0.0153)	(0.0697)	(0.0172)	(0.0071)	(0.0018)	(0.0093)	(0.0023)
500	8.0435	8.0476	8.0834	8.0874	12.1055	12.1116	12.2463	12.2524	0.4826	0.4829	0.6386	0.6389
	(0.0324)	(0.0081)	(0.0289)	(0.0072)	(0.0488)	(0.0121)	(0.0549)	(0.0137)	(0.0056)	(0.0014)	(0.0073)	(0.0018)

Table 5. *B.Es* and *P.Rs* under Gamma Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 4, p = 0.45$		$\alpha_1 = 4, p = 0.60$		$\alpha_2 = 6, p = 0.45$		$\alpha_2 = 6, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	4.3777	4.4002	4.2672	4.2891	6.5884	6.6222	6.5287	6.5623	0.5034	0.5060	0.6507	0.6541
	(0.1835)	(0.0418)	(0.1788)	(0.0407)	(0.2761)	(0.0629)	(0.2736)	(0.0623)	(0.0428)	(0.0097)	(0.0591)	(0.0135)
100	4.3451	4.3561	4.2354	4.2461	6.5393	6.5559	6.4801	6.4966	0.4997	0.5009	0.6459	0.6475
	(0.0890)	(0.0204)	(0.0868)	(0.0199)	(0.1339)	(0.0308)	(0.1327)	(0.0305)	(0.0208)	(0.0048)	(0.0287)	(0.0066)
200	4.2618	4.2671	4.1542	4.1594	6.4140	6.4221	6.3559	6.3639	0.4901	0.4907	0.6335	0.6343
	(0.0425)	(0.0100)	(0.0414)	(0.0097)	(0.0639)	(0.0150)	(0.0634)	(0.0148)	(0.0099)	(0.0023)	(0.0137)	(0.0032)
300	4.1425	4.1460	4.0379	4.0413	6.2345	6.2397	6.1780	6.1832	0.4764	0.4768	0.6158	0.6163
	(0.0267)	(0.0064)	(0.0260)	(0.0063)	(0.0402)	(0.0097)	(0.0398)	(0.0096)	(0.0062)	(0.0015)	(0.0086)	(0.0021)
400	4.0456	4.0482	3.9435	3.9460	6.0887	6.0925	6.0336	6.0373	0.4652	0.4655	0.6014	0.6018
	(0.0191)	(0.0047)	(0.0186)	(0.0046)	(0.0287)	(0.0071)	(0.0284)	(0.0070)	(0.0044)	(0.0011)	(0.0061)	(0.0015)
500	4.0185	4.0206	3.9171	3.9191	6.0479	6.0509	5.9932	5.9962	0.4621	0.4624	0.5974	0.5977
	(0.0150)	(0.0037)	(0.0147)	(0.0036)	(0.0226)	(0.0056)	(0.0224)	(0.0056)	(0.0035)	(0.0009)	(0.0048)	(0.0012)

Table 6. *B.Es* and *P.Rs* under Gamma Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 8, p = 0.45$		$\alpha_1 = 8, p = 0.60$		$\alpha_2 = 12, p = 0.45$		$\alpha_2 = 12, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	8.5853	8.6295	8.6278	8.6722	12.9209	12.9873	13.0711	13.1383	0.5151	0.5178	0.6816	0.6851
	(0.3598)	(0.0819)	(0.3206)	(0.0730)	(0.5415)	(0.1233)	(0.6092)	(0.1387)	(0.0532)	(0.0121)	(0.0696)	(0.0158)
100	8.5214	8.5430	8.5636	8.5853	12.8247	12.8572	12.9738	13.0067	0.5113	0.5126	0.6765	0.6782
	(0.1745)	(0.0401)	(0.1555)	(0.0357)	(0.2627)	(0.0603)	(0.2955)	(0.0679)	(0.0258)	(0.0059)	(0.0338)	(0.0078)
200	8.3581	8.3686	8.3994	8.4100	12.5789	12.5947	12.7251	12.7412	0.5015	0.5021	0.6636	0.6644
	(0.0833)	(0.0195)	(0.0742)	(0.0174)	(0.1254)	(0.0294)	(0.1411)	(0.0331)	(0.0123)	(0.0029)	(0.0161)	(0.0038)
300	8.1241	8.1309	8.1643	8.1712	12.2268	12.2370	12.3690	12.3793	0.4874	0.4879	0.6450	0.6455
	(0.0524)	(0.0126)	(0.0466)	(0.0112)	(0.0788)	(0.0190)	(0.0886)	(0.0214)	(0.0077)	(0.0019)	(0.0101)	(0.0024)
400	7.9341	7.9391	7.9734	7.9784	11.9409	11.9484	12.0797	12.0873	0.4760	0.4763	0.6299	0.6303
	(0.0374)	(0.0092)	(0.0333)	(0.0082)	(0.0563)	(0.0139)	(0.0633)	(0.0156)	(0.0055)	(0.0014)	(0.0072)	(0.0018)
500	7.8810	7.8850	7.9201	7.9240	11.8610	11.8669	11.9989	12.0049	0.4729	0.4731	0.6257	0.6260
	(0.0295)	(0.0073)	(0.0263)	(0.0065)	(0.0444)	(0.0110)	(0.0499)	(0.0124)	(0.0044)	(0.0011)	(0.0057)	(0.0014)

Table 7. *B.Es* and *P.Rs* under Chi Square Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 4, p = 0.45$		$\alpha_1 = 4, p = 0.60$		$\alpha_2 = 6, p = 0.45$		$\alpha_2 = 6, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	4.4228	4.4455	4.3111	4.3333	6.6563	6.6905	6.5961	6.6300	0.5086	0.5112	0.6574	0.6608
	(0.1876)	(0.0427)	(0.1828)	(0.0416)	(0.2823)	(0.0643)	(0.2797)	(0.0637)	(0.0489)	(0.0111)	(0.0676)	(0.0154)
100	4.3899	4.4010	4.2791	4.2899	6.6067	6.6235	6.5470	6.5635	0.5048	0.5061	0.6526	0.6542
	(0.0910)	(0.0209)	(0.0887)	(0.0204)	(0.1369)	(0.0314)	(0.1357)	(0.0312)	(0.0237)	(0.0055)	(0.0328)	(0.0075)
200	4.3057	4.3111	4.1970	4.2023	6.4801	6.4883	6.4215	6.4295	0.4952	0.4958	0.6400	0.6409
	(0.0434)	(0.0102)	(0.0423)	(0.0099)	(0.0654)	(0.0153)	(0.0648)	(0.0152)	(0.0113)	(0.0027)	(0.0156)	(0.0037)
300	4.1852	4.1887	4.0796	4.0830	6.2987	6.3040	6.2417	6.2469	0.4813	0.4817	0.6221	0.6227
	(0.0273)	(0.0066)	(0.0266)	(0.0064)	(0.0411)	(0.0099)	(0.0407)	(0.0098)	(0.0071)	(0.0017)	(0.0098)	(0.0024)
400	4.0873	4.0899	3.9842	3.9867	6.1514	6.1553	6.0958	6.0996	0.4700	0.4703	0.6076	0.6080
	(0.0195)	(0.0048)	(0.0190)	(0.0047)	(0.0293)	(0.0072)	(0.0291)	(0.0072)	(0.0051)	(0.0013)	(0.0070)	(0.0017)
500	4.0600	4.0620	3.9575	3.9595	6.1103	6.1133	6.0550	6.0580	0.4669	0.4671	0.6035	0.6038
	(0.0154)	(0.0038)	(0.0150)	(0.0037)	(0.0231)	(0.0058)	(0.0229)	(0.0057)	(0.0040)	(0.0010)	(0.0055)	(0.0014)

Table 8. *B.Es* and *P.Rs* under Chi Square Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 8, p = 0.45$		$\alpha_1 = 8, p = 0.60$		$\alpha_2 = 12, p = 0.45$		$\alpha_2 = 12, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	8.6738	8.7184	8.7168	8.7616	13.0541	13.1212	13.2059	13.2738	0.5204	0.5231	0.6886	0.6922
	(0.3678)	(0.0837)	(0.3278)	(0.0746)	(0.5536)	(0.1260)	(0.6227)	(0.1418)	(0.0608)	(0.0138)	(0.0795)	(0.0181)
100	8.6093	8.6311	8.6519	8.6738	12.9569	12.9898	13.1076	13.1408	0.5166	0.5179	0.6835	0.6852
	(0.1784)	(0.0410)	(0.1590)	(0.0365)	(0.2685)	(0.0617)	(0.3021)	(0.0694)	(0.0295)	(0.0068)	(0.0386)	(0.0089)
200	8.4442	8.4549	8.4860	8.4967	12.7086	12.7246	12.8563	12.8725	0.5067	0.5073	0.6704	0.6712
	(0.0852)	(0.0200)	(0.0759)	(0.0178)	(0.1282)	(0.0300)	(0.1442)	(0.0338)	(0.0141)	(0.0033)	(0.0184)	(0.0043)
300	8.2079	8.2147	8.2485	8.2554	12.3529	12.3632	12.4965	12.5070	0.4925	0.4929	0.6516	0.6522
	(0.0535)	(0.0129)	(0.0477)	(0.0115)	(0.0805)	(0.0194)	(0.0906)	(0.0218)	(0.0088)	(0.0021)	(0.0116)	(0.0028)
400	8.0159	8.0210	8.0556	8.0607	12.0640	12.0715	12.2043	12.2119	0.4810	0.4813	0.6364	0.6368
	(0.0382)	(0.0094)	(0.0341)	(0.0084)	(0.0575)	(0.0142)	(0.0647)	(0.0160)	(0.0063)	(0.0016)	(0.0083)	(0.0020)
500	7.9623	7.9663	8.0017	8.0057	11.9832	11.9892	12.1226	12.1287	0.4777	0.4780	0.6321	0.6325
	(0.0302)	(0.0075)	(0.0269)	(0.0067)	(0.0454)	(0.0113)	(0.0511)	(0.0127)	(0.0050)	(0.0012)	(0.0065)	(0.0016)

Table 9. *B.Es* and *P.Rs* under Exponential Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 4, p = 0.45$		$\alpha_1 = 4, p = 0.60$		$\alpha_2 = 6, p = 0.45$		$\alpha_2 = 6, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	4.4724	4.4954	4.3595	4.3820	6.7310	6.7656	6.6701	6.7044	0.5143	0.5170	0.6648	0.6682
	(0.1939)	(0.0441)	(0.1890)	(0.0430)	(0.2918)	(0.0664)	(0.2891)	(0.0658)	(0.0520)	(0.0118)	(0.0719)	(0.0164)
100	4.4391	4.4504	4.3271	4.3381	6.6809	6.6978	6.6204	6.6372	0.5105	0.5118	0.6599	0.6616
	(0.0940)	(0.0216)	(0.0917)	(0.0211)	(0.1415)	(0.0325)	(0.1403)	(0.0322)	(0.0252)	(0.0058)	(0.0349)	(0.0080)
200	4.3540	4.3595	4.2441	4.2495	6.5528	6.5611	6.4935	6.5017	0.5007	0.5013	0.6472	0.6480
	(0.0449)	(0.0105)	(0.0438)	(0.0103)	(0.0676)	(0.0158)	(0.0670)	(0.0157)	(0.0121)	(0.0028)	(0.0166)	(0.0039)
300	4.2322	4.2357	4.1253	4.1288	6.3694	6.3748	6.3118	6.3171	0.4867	0.4871	0.6291	0.6296
	(0.0282)	(0.0068)	(0.0275)	(0.0066)	(0.0425)	(0.0102)	(0.0421)	(0.0101)	(0.0076)	(0.0018)	(0.0105)	(0.0025)
400	4.1332	4.1358	4.0289	4.0314	6.2205	6.2244	6.1642	6.1680	0.4753	0.4756	0.6144	0.6148
	(0.0202)	(0.0050)	(0.0196)	(0.0048)	(0.0303)	(0.0075)	(0.0301)	(0.0074)	(0.0054)	(0.0013)	(0.0075)	(0.0018)
500	4.1055	4.1076	4.0019	4.0039	6.1788	6.1819	6.1229	6.1260	0.4721	0.4724	0.6103	0.6106
	(0.0159)	(0.0040)	(0.0155)	(0.0039)	(0.0239)	(0.0059)	(0.0237)	(0.0059)	(0.0043)	(0.0011)	(0.0059)	(0.0015)

Table 10. *B.Es* and *P.Rs* under Exponential Prior

n	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*	*SELF*	*PLF*
	$\alpha_1 = 8, p = 0.45$		$\alpha_1 = 8, p = 0.60$		$\alpha_2 = 12, p = 0.45$		$\alpha_2 = 12, p = 0.60$		$p = 0.45$		$p = 0.60$	
50	8.7712	8.8163	8.8146	8.8599	13.2006	13.2685	13.3541	13.4228	0.5263	0.5290	0.6964	0.6999
	(0.3802)	(0.0866)	(0.3388)	(0.0771)	(0.5722)	(0.1303)	(0.6437)	(0.1465)	(0.0646)	(0.0147)	(0.0846)	(0.0193)
100	8.7059	8.7280	8.7490	8.7712	13.1024	13.1356	13.2547	13.2883	0.5224	0.5237	0.6912	0.6929
	(0.1844)	(0.0424)	(0.1643)	(0.0377)	(0.2776)	(0.0637)	(0.3123)	(0.0717)	(0.0314)	(0.0072)	(0.0410)	(0.0094)
200	8.5390	8.5498	8.5813	8.5921	12.8512	12.8674	13.0006	13.0170	0.5123	0.5130	0.6779	0.6788
	(0.0880)	(0.0206)	(0.0785)	(0.0184)	(0.1325)	(0.0310)	(0.1491)	(0.0349)	(0.0150)	(0.0035)	(0.0196)	(0.0046)
300	8.3000	8.3070	8.3411	8.3481	12.4915	12.5020	12.6368	12.6473	0.4980	0.4984	0.6589	0.6595
	(0.0553)	(0.0133)	(0.0493)	(0.0119)	(0.0833)	(0.0201)	(0.0937)	(0.0226)	(0.0094)	(0.0023)	(0.0123)	(0.0030)
400	8.1059	8.1110	8.1460	8.1511	12.1994	12.2070	12.3412	12.3490	0.4864	0.4867	0.6435	0.6439
	(0.0395)	(0.0098)	(0.0352)	(0.0087)	(0.0595)	(0.0147)	(0.0669)	(0.0165)	(0.0067)	(0.0017)	(0.0088)	(0.0022)
500	8.0517	8.0557	8.0915	8.0956	12.1177	12.1238	12.2586	12.2648	0.4831	0.4833	0.6392	0.6396
	(0.0312)	(0.0077)	(0.0278)	(0.0069)	(0.0469)	(0.0117)	(0.0528)	(0.0131)	(0.0053)	(0.0013)	(0.0069)	(0.0017)

Table 11. 95% credible intervals under Uniform Prior

n	$\alpha_1 = 4$			$\alpha_2 = 6$			$p = 0.45$		
	LL	*UL*	*UL – LL*	*LL*	*UL*	*UL – LL*	*LL*	*UL*	*UL – LL*
50	3.4241	4.9498	1.5257	5.1532	7.4494	2.2962	0.3938	0.5692	0.1755
100	3.4882	4.9129	1.4248	5.2497	7.3939	2.1443	0.4011	0.5650	0.1638
200	3.5092	4.8187	1.3096	5.2813	7.2522	1.9709	0.4036	0.5542	0.1506
300	3.5391	4.6839	1.1448	5.3263	7.0492	1.7229	0.4070	0.5386	0.1317
400	3.6231	4.5743	0.9512	5.4528	6.8844	1.4316	0.4167	0.5260	0.1094
500	3.6817	4.5437	0.8620	5.5410	6.8383	1.2973	0.4234	0.5225	0.0991

Table 12. 95% credible intervals under Jeffreys Prior

n	$\alpha_1 = 4$			$\alpha_2 = 6$			$p = 0.45$		
	LL	*UL*	*UL – LL*	*LL*	*UL*	*UL – LL*	*LL*	*UL*	*UL – LL*
50	3.3898	4.9003	1.5104	5.1017	7.3749	2.2732	0.3898	0.5635	0.1737
100	3.4533	4.8638	1.4105	5.1972	7.3200	2.1228	0.3971	0.5593	0.1622
200	3.4741	4.7705	1.2965	5.2285	7.1797	1.9512	0.3995	0.5486	0.1491
300	3.5037	4.6370	1.1334	5.2730	6.9787	1.7057	0.4029	0.5333	0.1303
400	3.5869	4.5286	0.9417	5.3983	6.8155	1.4172	0.4125	0.5208	0.1083
500	3.6449	4.4983	0.8534	5.4856	6.7699	1.2843	0.4192	0.5173	0.0981

Table 13. 95% credible intervals under Gamma Prior

n	$\alpha_1 = 4$			$\alpha_2 = 6$			$p = 0.45$		
	LL	***UL***	***UL – LL***	***LL***	***UL***	***UL – LL***	***LL***	***UL***	***UL – LL***
50	3.3213	4.8013	1.4799	4.9986	7.2259	2.2273	0.3820	0.5521	0.1702
100	3.3835	4.7655	1.3820	5.0922	7.1721	2.0799	0.3891	0.5480	0.1589
200	3.4039	4.6742	1.2703	5.1228	7.0346	1.9118	0.3914	0.5375	0.1461
300	3.4329	4.5433	1.1105	5.1665	6.8377	1.6712	0.3948	0.5225	0.1277
400	3.5144	4.4371	0.9227	5.2892	6.6778	1.3886	0.4042	0.5103	0.1061
500	3.5713	4.4074	0.8361	5.3748	6.6331	1.2584	0.4107	0.5069	0.0962

Table 14. 95% credible intervals under Chi Square Prior

n	$\alpha_1 = 4$			$\alpha_2 = 6$			$p = 0.45$		
	LL	***UL***	***UL – LL***	***LL***	***UL***	***UL – LL***	***LL***	***UL***	***UL – LL***
50	3.3556	4.8508	1.4952	5.0501	7.3004	2.2503	0.3859	0.5578	0.1719
100	3.4184	4.8146	1.3963	5.1447	7.2460	2.1014	0.3931	0.5537	0.1606
200	3.4390	4.7224	1.2834	5.1757	7.1071	1.9315	0.3955	0.5431	0.1476
300	3.4683	4.5902	1.1219	5.2198	6.9082	1.6885	0.3989	0.5279	0.1290
400	3.5507	4.4828	0.9322	5.3438	6.7467	1.4029	0.4083	0.5155	0.1072
500	3.6081	4.4528	0.8447	5.4302	6.7015	1.2713	0.4149	0.5121	0.0971

Table 15. 95% credible intervals under Exponential Prior

n	$\alpha_1 = 4$			$\alpha_2 = 6$			$p = 0.45$		
	LL	***UL***	***UL – LL***	***LL***	***UL***	***UL – LL***	***LL***	***UL***	***UL – LL***
50	3.3932	4.9052	1.5120	5.1068	7.3823	2.2755	0.3902	0.5641	0.1739
100	3.4568	4.8687	1.4119	5.2024	7.3274	2.1250	0.3975	0.5599	0.1624
200	3.4776	4.7754	1.2978	5.2338	7.1869	1.9532	0.3999	0.5492	0.1492
300	3.5072	4.6417	1.1345	5.2783	6.9858	1.7074	0.4033	0.5338	0.1305
400	3.5905	4.5332	0.9426	5.4037	6.8224	1.4187	0.4129	0.5213	0.1084
500	3.6486	4.5028	0.8542	5.4911	6.7767	1.2856	0.4196	0.5178	0.0982

Table 16. 95% posterior predictive intervals under different priors

n	Limits	Uniform	Jeffreys	Gamma	Chi square	Exponential
	LL	2.4773	2.4037	2.2609	2.3316	2.4110
50	**UL**	17.9778	17.4438	16.4078	16.9206	17.4967
	UL – LL	15.5005	15.0401	14.1469	14.5890	15.0858
	LL	2.6190	2.5412	2.3903	2.4650	2.5489
100	**UL**	17.5793	17.0572	16.0442	16.5455	17.1089
	UL – LL	14.9603	14.5160	13.6539	14.0805	14.5600
	LL	2.6666	2.5874	2.4338	2.5098	2.5953
200	**UL**	16.5876	16.0950	15.1391	15.6121	16.1438
	UL – LL	13.9210	13.5075	12.7053	13.1023	13.5485
	LL	2.7354	2.6541	2.4965	2.5745	2.6622
300	**UL**	15.2334	14.7810	13.9031	14.3376	14.8258
	UL – LL	12.4981	12.1269	11.4066	11.7631	12.1636
	LL	2.9350	2.8478	2.6787	2.7624	2.8564
400	**UL**	14.1895	13.7681	12.9504	13.3551	13.8098
	UL – LL	11.2546	10.9203	10.2717	10.5927	10.9534
	LL	3.0797	2.9882	2.8108	2.8986	2.9973
500	**UL**	13.9065	13.4935	12.6921	13.0887	13.5344
	UL – LL	10.8268	10.5052	9.8813	10.1901	10.5371

The simulation study has been conducted under the assumption of different priors using two loss functions. The performance of different estimators has been compared in terms of posterior risks and rate of convergence. It has been observed that the estimated value of the parameter converges to the true value of the parameter by increasing the sample size. This pattern is similar under each prior and for every loss function. In cases of non-informative priors, the estimates under the Jeffreys prior provide better convergence, while among informative priors the gamma prior gives comparatively rapid convergence. On the other hand, use of squared error loss function results in faster convergence than precautionary loss function. It is interesting to note that for each combination of prior and loss function, the increased values of the parameters impose a negative impact on the convergence rate of the estimates. However, increasing the value of weight parameter has a positive effect on the convergence of the corresponding estimators but convergence rate of the weight estimator itself becomes slower. All the parameters have been overestimated in almost all the cases and the extent of

overestimation is greater for larger true parametric values. On the whole, it can be assessed that the estimates under gamma prior using precautionary loss function have the best convergence rate. Some prior elicitation technique may further strengthen this argument.

It is apparent from the above tables that the magnitude of posterior risks is indirectly proportional to sample size, while it is directly related to the true value of the parameter. This property holds for all estimators. The estimates under the Jeffreys prior have smaller risks than those under uniform prior. Similarly, among all informative priors, the estimates using gamma prior are associated with the minimum risks. In a comparison of informative and non-informative priors, informative priors perform better. This indicates the dominance of informative priors over non-informative priors. The performance of the estimates (representing the corresponding component) has been positively affected by increasing the values of the weight parameter, but at the cost of inflated risks for the parameter itself. It can also be observed that the estimates under precautionary loss function have smaller risks than those based on squared error loss function, irrespective of choice of prior and the true parametric values. Therefore, in terms of posterior risks, the estimates using gamma prior based on precautionary loss function provide the best point estimation.

In case of interval estimation (presented in Tables 11-16), the widths of 95% credible intervals decrease when increasing the sample size. The least amount of widths for credible intervals has been observed under gamma prior. In addition, the posterior predictions tend to be more specific under gamma prior. This is another indication that the gamma prior performs well as compared to other priors.

Real Life Example

Real life data regarding cancer survival times in years presented by Bekker et al. (2000) has been analyzed to illustrate the practical applicability of the results. The test termination time is considered to be such that the overall sample is 15% censored.

The analysis of real life data replicated the patterns observed under simulation study. The point and interval estimates for the parameters of the Gompertz mixture model; based on gamma prior; using *PLF* are found to be the most efficient. The posterior predictive intervals have the least amounts of widths again under gamma prior. So, in order to estimate the said parameters and to make predictions of the future values of the variable from the mentioned model, the use of gamma prior under *PLF* may be preferred.

Table 17. Bayes estimates and posterior risks under real life data

Priors	*SELF*			*PLF*		
	α_1	α_2	$p = 0.45$	α_1	α_2	$p = 0.45$
Uniform	2.0841	2.3456	0.4963	2.0948	2.3577	0.4988
	(1.1218)	(1.0466)	(0.1298)	(0.2554)	(0.2382)	(0.0296)
Jefrreys	2.0632	2.3222	0.4913	2.0738	2.3341	0.4939
	(1.1106)	(1.0361)	(0.1169)	(0.2528)	(0.2359)	(0.0266)
Gamma	2.0215	2.2753	0.4814	2.0319	2.2870	0.4839
	(1.0096)	(0.9419)	(0.0909)	(0.2298)	(0.2144)	(0.0207)
Chi Square	2.0424	2.2987	0.4864	2.0529	2.3105	0.4889
	(1.0321)	(0.9629)	(0.1039)	(0.2349)	(0.2192)	(0.0236)
Exponential	2.0653	2.3245	0.4918	2.0759	2.3365	0.4944
	(1.0668)	(0.9953)	(0.1105)	(0.2429)	(0.2266)	(0.0252)

Table 18. 95% credible intervals under real life data

Priors	α_1			α_2			$p = 0.45$		
	LL	*UL*	*UL–LL*	*LL*	*UL*	*UL–LL*	*LL*	*UL*	*UL–LL*
Uniform	1.7475	2.4206	0.6731	2.0316	2.8689	0.8373	0.3924	0.5872	0.1948
Jeffreys	1.7300	2.3964	0.6664	2.0113	2.8402	0.8289	0.3978	0.5731	0.1753
Gamma	1.7186	2.3244	0.6058	1.9927	2.7462	0.7535	0.4087	0.5450	0.1363
Chi Square	1.7328	2.3520	0.6192	2.0099	2.7801	0.7703	0.4033	0.5591	0.1558
Exponential	1.7452	2.3854	0.6401	2.0259	2.8222	0.7962	0.4034	0.5692	0.1657

Table 19. 95% posterior predictive intervals under real life data

Priors	*LL*	*UL*	*UL–LL*
Uniform	0.6227	11.0830	10.4602
Jeffreys	0.6165	10.9721	10.3556
Gamma	0.6124	10.6426	10.0302
Chi Square	0.6175	10.7688	10.1514
Exponential	0.6219	10.9216	10.2997

Conclusion

The study proposed the point and interval estimators for the parameters of the two-component mixture of the Gompertz distribution under a Bayesian framework along with posterior predictions for the future value from said model. The performance of the different estimators has been compared in terms of posterior risks (for point estimators) and widths of interval estimates with respect to various priors and loss functions. The findings of the study suggest that for Bayesian estimation of the parameters (along with posterior predictions) of the two-component mixture of the Gompertz distribution, the use of gamma prior under precautionary loss function is preferred. The proposed estimators are consistent in nature. The results of the study are useful for practitioners looking to model some failure time data, where the cases of failures are more than one.

References

Aslam, M. (2003). An application of prior predictive distribution to elicit the prior density. *Journal of Statistical Theory and Applications*, *2*(1), 183-197.

Bekker, A., Roux, J., & Mostert, P. (2000). A generalization of the compound Rayleigh distribution: using a Bayesian method on cancer survival times. *Communication in Statistics-Theory and Methods*, *29*(7), 1419-1433.

Eberlya, L. E., & Casella, G. (2003). Estimating Bayesian credible intervals. *Journal of Statistical Planning and Inference*, *112*, 115-132.

Ismail, A. A. (2010). Bayes estimation of Gompertz distribution parameters and acceleration factor under partially accelerated life tests with type-I censoring. *Journal of Statistical Computation and Simulation*, *80*(11), 1253-1264.

Ismail, A. A. (2011). Planning step-stress life tests with type-II censored Data. *Scientific Research and Essays*, *6*(19), 4021-4028.

Jeffreys, H. (1961). *Theory of Probability*. 3rd Ed. Oxford: Oxford University Press.

Jaheen, Z. F. (2003). A Bayesian analysis of record statistics from the Gompertz model. *Applied Mathematical Computations*, *145*(2-3), 307-320.

Kazmi, S.M.A., Aslam, M., & Ali, S. (2012). On the Bayesian estimation for two component mixture of maxwell distribution, assuming type I censored data. *International Journal of Applied Science and Technology (IJAST)*, *2*(1), 197-218.

Khedhairi, A., & Gohary, A. E. (2008). A new class of bivariate Gompertz distributions and its mixture. *International Journal of Mathematical Analysis, 2*(5), 235-253.

Kiani, K., Arasan, J., & Midi, H. (2012). Interval estimations for parameters of Gompertz model with time-dependent covariate and right censored data. *Sains Malaysiana, 41*(4), 471-480.

Laplace, P. S. (1812). *Theorie Analytique Des Probabilities*. Veuve Courcier Paris.

Majeed, M.Y., & Aslam, M. (2012). Bayesian analysis of the two component mixture of inverted exponential distribution under quadratic loss functions. *International Journal of Physical Sciences,* 7(9), 1424-1434.

Pollard, J., & Valkovics, E. (1992). The Gompertz distribution and its applications. *Genus, 48*(3-4), 15-29.

Saleem, M., & Aslam, M. (2008). Bayesian analysis of the two component mixture of the Rayleigh dist. With the uniform and the Jeffreys priors. *Journal of Applied Statistical Science, 16*(4), 105-113.

Saleem, M., Aslam, M., & Economou, P. (2010). On the Bayesian analysis of the mixture of power function distribution using the complete and the censored sample. *Journal of Applied Statistics, 37*(1), 25-40.

Saracoglu, B., Kaya, M.F. & Abd-Elfattah, A.M. (2009). Comparison of estimators for stress-strength reliability in the Gompertz case. *Hacettepe Journal of Mathematics and Statistics, 38*(*3*), 339-349.

Willemse, W. and Koppelaar, H. (2000). Knowledge elicitation of Gompertz' law of morality, *Scandinavian Actuarial Journal, 94*, 168-180.

Wu, C. C., Wu, S. F., & Chan, H. Y. (2006). MLE and the estimated expected test time for the two-parameter Gompertz distribution under progressive censoring with binomial removals. *Applied Mathematical Computation, 181*(2), 1657-1670.

Wu, J.W., Hung, W.L., & Tsai, C.H. (2004). Estimation of parameters of the Gompertz distribution using the least squares method. *Applied Mathematical Computation, 158*(1), 133-147.

Wu, S.J., Chang, C.T., & Tsai, T.R. (2003). Point and interval estimations for the Gompertz distribution under progressive type-II censoring. *Metron - International Journal of Statistics, 21*(3), 403-418.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 293-303.

ISSN 1538 − 9472

A Comparison between Biased and Unbiased Estimators in Ordinary Least Squares Regression

Ghadban Khalaf
King Khalid University
Saudi Arabia

During the past years, different kinds of estimators have been proposed as alternatives to the Ordinary Least Squares (*OLS*) estimator for the estimation of the regression coefficients in the presence of multicollinearity. In the general linear regression model, $\vec{Y} = X\vec{\beta} + \vec{e}$, it is known that multicollinearity makes statistical inference difficult and may even seriously distort the inference. Ridge regression, as viewed here, defines a class of estimators of $\vec{\beta}$ indexed by a scalar parameter *k*. Two methods of specifying *k* are proposed and evaluated in terms of Mean Square Error (*MSE*) by simulation techniques. A comparison is made with other ridge-type estimators evaluated elsewhere. The estimated *MSE* of the suggested estimators are lower than other estimators of the ridge parameter and the *OLS* estimator.

Keywords: *OLS* estimator, linear regression, multicollinearity, ridge regression, Monte Carlo simulation.

Introduction

Consider the multiple linear regression model

$$\vec{Y} = X\vec{\beta} + \vec{e} \tag{1}$$

where $\vec{Y}$ is an $(n\times 1)$ response vector, X is a fixed $(n\times p)$ matrix of independent variables of rank p, $\vec{\beta}$ is the unknown $(p\times 1)$ parameter vector of regression coefficients and, finally, $\vec{e}$ is an $(n\times 1)$ vector of uncorrelated errors with mean zero and common unknown variance σ^2. If $X'X$ is nonsingular, the *OLS* estimator for $\vec{\beta}$ is given by

Ghadban Khalaf is an Associate Professor in the Department of Mathematics.

$$\hat{\beta} = (X'X)^{-1} X'\vec{Y} \tag{2}$$

For orthogonal data, the *OLS* estimator in the linear regression model is strongly efficient. But in the presence of multicollinearity, the *OLS* efficiency can be reduced and hence an improvement upon it would be necessary and desirable.

The term multicollinearity is used to denote the presence of linear relationships, or near linear relationships, among explanatory variables. If the explanatory variables are perfectly linearly correlated, that is, if the correlation coefficient for these variables is equal to unity, then the parameters become indeterminate; i.e, it is impossible to obtain numerical values for each parameter separately and the method of lease squares breaks down. Conversely, if the correlation coefficient for the explanatory variables is equal to zero, then the variables are called orthogonal and there are no problems concerning the estimates of the coefficients.

When multicollinearity occurs, the least squares estimates are still unbiased and efficient but the problem is that; the estimated standard error $S_{\hat{\beta}_i}$ for the coefficient $\hat{\beta}_i$ become infinitely large; i.e, the standard error tends to be larger than it would be in the absence of multicollinearity and when $S_{\hat{\beta}_i}$ is larger than it would be, then the *t*- value for testing the significance of β_i is smaller than it should be. Thus one is likely to conclude that a variable X_i is not important in the relationship when it real1y is.

To solve the problem of multicollinearity, there is no single solution that will eliminate multicollinearity altogether. One common procedure is to select the independent variable most seriously involved in the multicollinearity and remove it from the model. This procedure often improves the standard error of the remaining coefficients and may make formerly insignificant variables significant, since the elimination of a variable reduces any multicollinearity caused by it. The difficulty with this approach is that the model now may not correctly represent the population relationship and all estimated coefficients would contain a population specification.

The procedure of increasing the sample size is sometimes recommended as another suggested procedure to solve the problem of multicollinearity. In fact this method improves the precision of an estimator and hence reduces the adverse effects of multicollinearity.

Hoerl and Kennard (1970) suggested a new technique to overcome the problem of multicollinearity. This technique is called ridge regression. Ridge

regression is a variant of ordinary multiple linear regression whose goal is to circumvent the predictors collinearity. It gives up the least squares as a method for estimating the parameters of the model and focuses instead of the $X'X$ matrix. This matrix will be artificially modified so as to make its determinant appreciably different from zero. This is accomplished by adding a small positive quantity, say k ($k > 0$), to each of the diagonal elements of the matrix $X'X$ before inverting it for least squares estimation. The resulting estimator is given by

$$\vec{\hat{\beta}}(k) = (X'X + kI_p)^{-1} X'\vec{Y}, \quad k > 0 \tag{3}$$

which coincides with the *OLS* estimator, defined by (2), when $k = 0$. The resulting estimator will be biased, but have smaller variances than $\vec{\hat{\beta}}$. This is precisely what the ridge regression estimator we study can accomplish.

The plan of this paper is as follows: Section 2 presents the proposed estimators included in the study; a novel feature is our proposed ridge estimator which, as we shall see presently, has lower *MSE*. Section 3 is described the simulation technique that we have adopted in our study to evaluate the performance of the new values of the ridge parameter we suggest. The results of the simulation study, which appear in the tables, are presented in Section 4. Finally, Section 5 contains summary and conclusions.

The Proposed Estimators

With the ridge estimator method, there arises the problem of determining an optimal value of k. With a good choice of k, one might hope to improve on the *OLS* estimator for every coefficient.

Hoerl, Kennard and Baldwin (1975) showed, through simulation, that the use of ridge estimator with the following biasing parameter

$$\hat{k} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \hat{\beta}_i^2} \tag{4}$$

implies that $MSE(\vec{\hat{\beta}}(k)) < MSE(\vec{\hat{\beta}})$, where p denotes the number of parameters (excluding the intercept) and $\hat{\sigma}^2$ is the usual unbiased estimate of σ^2, defined by;

$$\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2 / (n - p - 1).$$

They showed that the probability of a smaller *MSE* using (4) increases with the number of parameters *p*. We will use the acronym *HKB* for the estimator (4).

Khalaf and Shukur (2005) suggested a modification of Hoerl and Kennard (1970) given by;

$$\hat{k} = \frac{t_{\max}\,\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_{\max}\hat{\beta}^2_{\max}} \tag{5}$$

which guaranteed lower *MSE*, where $t_{\max}$ is the maximum eigenvalue of $X'X$ matrix. For this estimator we will use the acronym *KS*.

From the estimators (4) and (5), we suggest as a modification of *HKB* and *KS* by multiplying them by the amount;

$$\frac{\frac{1}{2}(t_{\max} + t_{\min})}{\sum_{i=1}^{p}\left|\hat{\beta}_i\right|} = \frac{t_{\max} + t_{\min}}{2\sum_{i=1}^{p}\left|\hat{\beta}_i\right|},$$

where $t_{\min}$ is the minimum eigenvalue of the matrix $X'X$. This leads to the following estimators;

$$\hat{k}_1 = \frac{(t_{\max} + t_{\min})}{2\sum_{i=1}^{p}\left|\hat{\beta}_i\right|} . \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\hat{\beta}_i^2} \tag{6}$$

$$\hat{k}_2 = \frac{(t_{\max} + t_{\min})\, t_{\max}\,\hat{\sigma}^2}{2\sum_{i=1}^{p}\left|\hat{\beta}_i\right|((n-p)\hat{\sigma}^2 + t_{\max}\hat{\beta}^2_{\max})}. \tag{7}$$

For our two suggested estimators, defined by (6) and (7), we use the acronym K_1 and K_2, respectively.

The Simulation Study

A simulation study was conducted in order to draw conclusions about the performance of our suggested estimators relative to *HKB, KS* and the *OLS* estimator. To achieve different degree of collinearity, following Kibria (2003), the explanatory variables are generated by using the following equation;

$$x_{ij} = (1-\rho^2)^{1/2} z_{ij} + \rho\, z_{ip}, \qquad i = 1,2,...,n. \qquad j = 1,2,...,p$$

where z_{ij} are independent standard normal distribution, p is the number of the explanatory variables and ρ is specified so that the correlation between any two explanatory variables is given by ρ^2. Three different sets of correlation are considered according to the value of ρ = 0.85, 0.95 and 0.99. The n observations for the dependent variable are determined by the following equation;

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + e_i, \qquad i = 1,2,...,n$$

where e_i are *i.i.d* pseudo-random numbers. In this study, β_0 is taken to be zero and the term e_i is generated from each of the following distributions: $N(0, 1)$, $T(3)$, $T(7)$ and $F(3, 11)$. The parameters values are chosen so that $\sum_{i=1}^{p} \beta_i^2 = 1$, which is a common restriction in simulation studies (see Muniz and Kibria (2009)).

The other factors we chose to vary is the sample size and the number of regressions. We generate models consisting of 25, 50, 100 and 150 observations and with 2 and 4 explanatory variables. It is noted from the results of the previous simulation studies (see Khalaf and Shukur (2005), Alkhamisi and Shukur (2008) and Khalaf (2011)) that increasing the number of regressor and using non-normal pseudo random numbers to generate e_i leads to a higher estimated *MSE*, while increasing the sample size leads to a lower estimated *MSE*.

The criterion proposed here for measuring the goodness of an estimator is the *MSE* using the following formula;

$$MSE(\vec{\hat{\beta}}_r) = \frac{1}{5000} \sum (\vec{\hat{\beta}}_r - \vec{\beta})'(\vec{\hat{\beta}}_r - \vec{\beta}), \qquad (8)$$

where $\vec{\hat{\beta}}_r$ is the estimator of $\vec{\beta}$ obtained from the *OLS* estimator or from the ridge estimator for different estimated values of *k* considered for comparison reasons and, finally, 5000 is the number of replicates used in the Monte Carlo simulation.

The Simulation Results

Tables 1 – 6 below, present the output from the Monte Carlo experiment concerning properties of the different methods that used to choose the ridge parameter *k*. The results showed that the estimated *MSE* is affected by all factors we choose to vary in the design of experiment. It is also noted that the higher the degree of correlation the higher estimated *MSE*, but this increase is much greater for the *OLS* than the ridge regression estimator. The distribution of the error term and the number of explanatory variables having a different impact on the estimators.

Table 1. Estimated *MSE* when $\rho = 0.85$ and $p = 2$.

	OLS	***HKB***	***KS***	K_1	K_2
N(0, 1)					
25	0.238	0.181	0.190	0.243	0.181
50	0.111	0.093	0.097	0.255	0.176
100	0.057	0.051	0.053	0.266	0.178
150	0.034	0.032	0.032	0.282	0.184
T(3)					
25	2.259	0.957	1.325	0.506	0.497
50	1.219	0.602	0.896	0.521	0.364
100	0.531	0.312	0.445	0.588	0.329
150	0.473	0.261	0.414	0.632	0.351
T(7)					
25	0.350	0.248	0.266	0.304	0.218
50	0.169	0.135	0.143	0.324	0.212
100	0.076	0.067	0.069	0.352	0.220
150	0.053	0.048	0.049	0.367	0.224
F(3, 11)					
25	0.853	0.502	0.584	0.383	0.289
50	0.391	0.261	0.309	0.429	0.236
100	0.178	0.139	0.157	0.481	0.276
150	0.126	0.104	0.115	0.520	0.290

Table 2. Estimated *MSE* when $\rho = 0.85$ and $p = 4$.

	OLS	*HKB*	*KS*	K_1	K_2
N(0, 1)					
25	0.796	0.549	0.572	0.187	0.159
50	0.334	0.255	0.268	0.216	0.162
100	0.156	0.131	0.137	0.257	0.182
150	0.103	0.090	0.093	0.284	0.196
T(3)					
25	7.387	3.781	4.330	1.049	1.205
50	6.222	2.961	4.179	1.073	1.622
100	1.685	0.969	1.318	0.509	0.409
150	1.240	0.754	1.018	0.551	0.332
T(7)					
25	1.159	0.730	0.776	0.212	0.178
50	0.504	0.362	0.389	0.255	0.184
100	0.235	0.188	0.200	0.323	0.218
150	0.152	0.127	0.135	0.353	0.231
F(3, 11)					
25	2.667	1.446	1.601	0.338	0.328
50	1.130	0.699	0.805	0.316	0.226
100	0.578	0.402	0.468	0.415	0.263
150	0.362	0.271	0.311	0.467	0.282

Table 3. Estimated MSE when $\rho = 0.95$ and p = 2.

	OLS	*HKB*	*KS*	K_1	K_2
N(0, 1)					
25	0.705	0.427	0.450	0.193	0.147
50	0.353	0.250	0.265	0.193	0.134
100	0.168	0.133	0.140	0.220	0.145
150	0.114	0.095	0.010	0.231	0.149
T(3)					
25	7.899	3.010	3.580	1.501	1.725
50	5.575	2.137	2.969	0.988	1.256
100	1.703	0.789	1.152	0.486	0.275
150	1.283	0.655	0.959	0.541	0.299
T(7)					
25	1.174	0.670	0.718	0.241	0.186
50	0.528	0.340	0.371	0.249	0.164
100	0.250	0.185	0.200	0.287	0.177
150	0.161	0.127	0.137	0.311	0.187
F(3, 11)					
25	2.556	1.223	1.372	0.401	0.343
50	1.167	0.623	0.738	0.336	0.215
100	0.566	0.346	0.419	0.397	0.224
150	0.378	0.251	0.300	0.444	0.244

Table 4. Estimated *MSE* when $\rho = 0.95$ and $p = 4$.

	OLS	*HKB*	*KS*	K_1	K_2
N(0, 1)					
25	2.356	1.308	1.351	0.154	0.145
50	1.057	0.673	0.705	0.125	0.099
100	0.359	0.251	0.266	0.120	0.088
150	0.323	0.245	0.258	0.194	0.138
T(3)					
25	18.886	8.484	9.026	2.145	2.299
50	14.573	7.274	8.218	1.939	2.315
100	4.122	2.079	2.527	0.284	0.194
150	3.390	1.815	2.266	0.422	0.306
T(7)					
25	3.716	1.996	2.068	0.228	0.221
50	1.502	0.892	0.948	0.146	0.114
100	0.745	0.495	0.534	0.199	0.139
150	0.478	0.341	0.368	0.239	0.162
F(3, 11)					
25	8.220	4.148	4.334	0.776	0.796
50	3.578	1.882	2.064	0.206	0.171
100	1.755	1.034	1.170	0.248	0.164
150	1.180	0.741	0.849	0.309	0.195

Table 5. Estimated *MSE* when $\rho = 0.99$ and $p = 2$.

	OLS	*HKB*	*KS*	K_1	K_2
N(0, 1)					
25	4.050	1.850	1.905	0.349	0.331
50	1.776	0.884	0.931	0.133	0.099
100	0.913	0.533	0.568	0.138	0.091
150	0.572	0.358	0.385	0.155	0.099
T(3)					
25	43.786	15.618	16.407	12.046	12.512
50	21.736	7.673	8.510	3.155	3.377
100	8.794	3.602	4.217	0.481	0.399
150	7.046	2.461	3.274	0.362	0.231
T(7)					
25	6.108	2.657	2.745	0.561	0.544
50	2.623	1.192	1.274	0.171	0.124
100	1.370	0.732	0.797	0.178	0.111
150	0.865	0.502	0.551	0.204	0.123
F(3, 11)					
25	12.863	4.822	5.037	1.421	1.438
50	6.402	2.550	2.779	0.343	0.296
100	3.329	1.508	1.715	0.246	0.152
150	1.901	0.899	1.060	0.279	0.151

Table 6. Estimated *MSE* when $\rho = 0.99$ and $p = 4$.

	OLS	*HKB*	*KS*	K_1	K_2
N(0, 1)					
25	13.319	6.484	6.547	0.971	0.981
50	6.095	3.078	3.141	0.109	0.107
100	2.708	1.466	1.520	0.061	0.050
150	1.720	0.990	1.036	0.076	0.057
T(3)					
25	169.385	72.238	73.397	58.482	59.442
50	65.170	33.982	34.732	15.466	15.685
100	30.913	15.077	15.913	2.328	2.448
150	19.922	8.885	9.738	0.505	0.556
T(7)					
25	19.789	9.337	9.473	1.739	1.756
50	8.782	4.342	4.442	0.230	0.229
100	4.068	2.152	2.240	0.077	0.063
150	2.550	1.390	1.467	0.086	0.062
F(3, 11)					
25	44.422	20.834	21.062	7.010	7.089
50	21.347	9.485	9.785	1.073	1.131
100	9.172	4.498	4.730	0.145	0.131
150	6.178	3.077	3.303	0.114	0.085

For non-normal error term in combination with $\rho = 0.95$ and $\rho = 0.99$ leads to a larger estimated *MSE* for the *OLS* estimator and the ridge parameter, especially when *n* is small, but when the sample size increases the estimated *MSE* of the suggested ridge parameters, namely K_1 and K_2 decreases substantially.

The performance of K_1 and K_2 is well for all cases when the error term is distributed as a normal and, when *n* is greater than 25 and the error term in non-normal .

When *n* is greater than 25, the modified ridge parameter performance, defined by (6) and (7), is much better than the estimators *HKB*, *KS* and the *OLS*, where K_2 has a low estimated *MSE* when the number of regressor equals 4.

Summary and Conclusions

In multiple linear regression, the effect of non-orthogonality of the explanatory variables is to pull the least squares estimates of the regression coefficients away from the true coefficients, $\vec{\beta}$, that one is trying to estimate. The coefficients can

be both too large in absolute value and incorrect with respect to sign. Furthermore, the variance and the covariance of the *OLS* tend to become too large.

A slight movement away from this point can give completely different estimates of the coefficients. This is accomplished by adding a small positive quantity, *k*, to each of the diagonal elements of the matrix $X'X$. The resulting estimator is called the ridge estimator, suggested by Hoerl and Kennard (1970) and given by (3).

Several procedure for constructing ridge estimators have been proposed in the literature. These procedures were aiming at a rule (or algorithm) for selecting the constant *k* in equation (3). In fact, the best method of estimating *k* is an unsolved problem and there is no constant value of *k* that is certain to yield an estimator which is uniformly better (in terms of *MSE*) than the *OLS* in all cases.

By means of Monte Carlo simulations two suggested ridge parameters were evaluated and the result were compared with ridge parameters evaluated by Hoerl et. al (1975) and Khalaf and Shukur (2005). The estimator *HKB* performed well in this study. It appears to outperform *KS* when ρ is small and the sample size is greater than 25. The suggested estimators K_1 and K_2 performs well in our simulation. They appeared to offer an opportunity for large reduction in *MSE* when $p = 2$ and the error term in normally distributed. For non-normal error term the versions of the ridge parameter has a lower estimated *MSE* when the sample size is greater than 25. K_2 is always minimizes the estimated *MSE* when the error term in normally distributed.

References

Alkhamisi, M., & Shukur, G. (2008). Developing Ridge Parameters for SUR Model. *Communication in Statistics- Theory and Methods, 37*, 544-564.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for non-orthogonal Problems. *Technometrics, Vol. 12*, 55 – 67.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge Regression: some Simulation. *Communications in Statistics- Theory and Methods, 4*, 105 – 124.

Khalaf, G. (2011). Suggested Ridge Regression Estimators Under Multicollinearity. *Journal of Natural and Applied Sciences, University of Aden, Vol. 15*, 170 – 193.

Khalaf, G., & Shukur, G. (2005). Choosing Ridge Parameters for Regression Problems. *Communication in Statistics – Theory and Methods, 34,* 1177 – 1182.

Kibria, B. M. G. (2003). Performance of some New Ridge Regression Estimators. *Communication in Statistics- Theory and Methods, 32*, 419-435.

Muniz, G., & Kibria, B. M. G. (2009). On some Ridge Regression Estimators: An Empirical Comparisons. *Communications in Statistics-Simulation and Computation, 38,* 621 – 630.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 304-313.

ISSN 1538 − 9472

Comparison of Parameters of Lognormal Distribution Based On the Classical & Posterior Estimates

Raja Sultan
University of Kashmir
Srinagar, India

S. P. Ahmad
University of Kashmir
Srinagar, India

Lognormal distribution is widely used in scientific field, such as agricultural, entomological, biology etc. If a variable can be thought as the multiplicative product of some positive independent random variables, then it could be modelled as lognormal. In this study, maximum likelihood estimates and posterior estimates of the parameters of lognormal distribution are obtained and using these estimates we calculate the point estimates of mean and variance for making comparisons.

Keywords: Lognormal distribution, maximum likelihood estimation, posterior estimates & R software

Introduction

Aitchison & Brown (1957) have given a very comprehensive treatment of lognormal distribution. The lognormal distribution arises in various different contexts such as in physics (distribution of particles due to pulverisation); economics (income distributions); biology (growth of organisms), etc. Epstein (1947), Brownlee (1949), Delaporte (1950), Moroney (1951) describes various applications of lognormal distribution to physical and industrial processes, textile research and quality control. In the context of life testing and reliability problems, the lognormal distribution answers a criticism sometimes raised against the use of normal distribution (ranging from -∞ to +∞) as a model for the failure time distribution which must range from 0 to ∞.

A random variable X is said to have a lognormal distribution if $U = \log_e X$ has normal distribution with mean μ and variance σ^2. Thus, the pdf of lognormal distribution is given by

Dr. Sultan is a research scholar in the Department of Statistics. Email him at: hamzasultan18@yahoo.com. Dr. Ahmad is an Assistant Professor of in the Department of Statistics. Email him at: sprvz@yahoo.com.

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma x}\exp(-\frac{1}{2\sigma^2}(\log x-\mu)^2) \quad ,-\infty<\mu<\infty,\ \sigma^2>0,\ \ 0\le x<\infty \qquad (1)$$

The likelihood function of the random sample $(x_1, x_2, x_3, ..., x_n)^T$ would be

$$L(\mu,\sigma^2\mid x)=\left(\frac{1}{\sqrt{2\pi}\,\sigma x}\right)^n \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i-\mu)^2) \qquad (2)$$

The mean and variance of the lognormal distribution are given by

$$E(X)=\alpha_1=\exp\left(\mu+\frac{\sigma^2}{2}\right) \qquad (3)$$

and

$$V(X)=\beta_1=\exp\left(2\mu+\sigma^2\right)\left(\exp(\sigma^2)-1\right) \qquad (4)$$

Maximum Likelihood Estimators

Maximum Likelihood is a popular estimation technique for many distributions because it picks the values of the distribution's parameters that make the data “more likely" than any other values of the parameters would make them. This is accomplished by maximizing the likelihood function of the parameters given the data.

Consider the estimation of the parameters α1 and β1. Let $U_i=\log x_i$, $i=1,2,\ldots,n$. Then using the fact that $(U_1, U_2 , ... , U_n)$ is a random sample from Normal distribution with parameters (μ,σ^2). The mle of μ and σ^2 first are given by

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}U_i=\bar{U} \qquad (5)$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(U_i - \bar{U})^2 \tag{6}$$

The mle of α_1 and β_1 are given by

$$\hat{\alpha}_1 = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) \tag{7}$$

and

$$\hat{\beta}_1 = \exp\left(2\hat{\mu} + \hat{\sigma}^2\right)\left(\exp\left(\hat{\sigma}^2\right) - 1\right) \tag{8}$$

Posterior estimation of the parameter

Again, consider the estimation of the parameters α_1 and β_1. First obtain the posterior estimates of μ and σ^2 and then simultaneously the posterior estimates for α_1 and β_1 will be obtained. Laplace (1774) found that it worked exceptionally well to simply always choose the prior probability distribution for the parameter(s) of the model to be constant on the parameter space.

The joint prior pdf for μ and σ^2 considered is

$$P\left(\mu, \sigma^2\right) \propto 1 \tag{9}$$

According to Bayes theorem, Joint posterior density of μ and σ^2 is given by

Posterior density $\propto$ prior density* likelihood

$$\pi\left(\mu, \sigma^2 \mid \underset{\sim}{x}\right) \propto P\left(\mu, \sigma^2\right).P\left(\mu, \sigma^2 \mid \underset{\sim}{x}\right)$$

From equation (2) and (9) the joint posterior density of μ and σ^2 is given by

$$\pi\left(\mu, \sigma^2 \mid \underset{\sim}{x}\right) \propto \left[\frac{1}{\sqrt{2\pi\sigma x}}\right]^n \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(\log x_i - \mu\right)^2\right\}$$

$$\pi\left(\mu,\sigma^2 \mid \underset{\sim}{x}\right)=\frac{c}{\left(\sigma^2\right)}\exp\left(\frac{-\beta}{2\sigma^2}\right)\exp\left\{\frac{-n}{2\sigma^2}\left(\mu-\frac{\sum_{i=1}^{n}\log x_i}{n}\right)^2\right\} \qquad (10)$$

where $\beta=\sum_{i=1}^{n}\left(\log x_i\right)^2-\frac{\left(\sum_{i=1}^{n}\log x_i\right)^2}{n}$ and c is a normalizing constant. Lindley (1961) explained if $P(\theta)$ be the prior and $P\left(\underset{\sim}{x}\mid\theta\right)$ be the likelihood, the posterior pdf $P\left(\theta\mid\underset{\sim}{x}\right)$ is given by $P\left(\theta\mid\underset{\sim}{x}\right)=cP(\theta).P\left(\underset{\sim}{x}\mid\theta\right)$, where c is the normalizing constant. Then the value of c is obtained by $c=\left[\int P(\theta).P\left(\underset{\sim}{x}\mid\theta\right)d\theta\right]^{-1}$

Therefore, c can be obtained by

$$c^{-1}=\int_{0}^{\infty}\int_{-\infty}^{\infty}\pi\left(\mu,\sigma^2\mid\underset{\sim}{x}\right)d\mu\,d\sigma^2$$

$$c^{-1}=\int_{o}^{\infty}\int_{-\infty}^{\infty}\frac{1}{\left(\sigma^2\right)^{n/2}}\exp\left(\frac{-\beta}{2\sigma^2}\right)\exp\left\{\frac{-n}{2\sigma^2}\left(\mu-\frac{\sum_{i=1}^{n}\log x_i}{n}\right)^2\right\}d\mu\;d\sigma^2$$

Using the transformation $t=\frac{\sqrt{n}\left(\mu-\frac{\sum_{i=1}^{n}\log x_i}{n}\right)}{\sigma}$

$$c^{-1} = \sqrt{\frac{2\pi}{n}} \int_{o}^{\infty} \frac{\exp\left(\frac{-\beta}{2\sigma^2}\right)}{\left(\sigma^2\right)^{n-1/2}} d\sigma^2$$

$$c^{-1} = \sqrt{\frac{2\pi}{n}} \frac{\Gamma\left(\frac{n-3}{2}\right)}{\left(\frac{\beta}{2}\right)^{\frac{n-3}{2}}}$$

$$\Rightarrow c = \sqrt{\frac{n}{2\pi}} \frac{(\beta)^{\frac{n-3}{2}}}{2^{\frac{n-3}{2}} \Gamma\left(\frac{n-3}{2}\right)} \tag{11}$$

From the equation (10)

$$\pi\left(\mu, \sigma^2 \mid \underset{-}{x}\right) = \sqrt{\frac{n}{2\pi}} \frac{(\beta)^{\frac{n-3}{2}}}{2^{\frac{n-3}{2}} \Gamma\left(\frac{n-3}{2}\right)\left(\sigma^2\right)^{n/2}} \exp\left(\frac{-\beta}{2\sigma^2}\right) \exp\left\{\frac{-n}{2\sigma^2}\left(\mu - \frac{\sum_{i=1}^{n} \log x_i}{n}\right)^2\right\} \tag{12}$$

Marginal posterior densities of μ and σ^2

The marginal density of μ is obtained by integrating out σ^2 from (12) and is given as

$$\pi\left(\mu \mid \underset{-}{x}\right) = \int_{0}^{\infty} \pi\left(\mu, \sigma^2 \mid \underset{-}{x}\right) d\sigma^2$$

$$\pi\left(\mu \mid \underset{\sim}{x}\right) = c\int_0^{\infty} \frac{1}{\left(\sigma^2\right)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left\{\beta + n\left(\mu - \frac{\sum_{i=1}^{n} \log x_i}{n}\right)^2\right\}\right] d\sigma^2$$

$$\pi\left(\mu \mid \underset{\sim}{x}\right) = c\frac{\Gamma\left(\frac{n}{2}-1\right)2^{\frac{n}{2}-1}}{\left[\beta + n\left(\mu - \frac{\sum_{i=1}^{n} \log x_i}{n}\right)^2\right]^{\frac{n}{2}-1}}$$

$$\pi\left(\mu \mid \underset{\sim}{x}\right) = \sqrt{\frac{n}{\beta}} \frac{1}{B\left(\frac{1}{2}, \frac{n-3}{2}\right)\left[1 + \frac{n}{\beta}\left(\mu - \frac{\sum_{i=1}^{n} \log x_i}{n}\right)^2\right]^{\frac{n-2}{2}}} \tag{13}$$

The marginal density of σ^2 is obtained by integrating the joint posterior density of μ and σ^2 given in (12) over the range of μ. It is given as

$$\pi\left(\sigma^2 \mid \underset{\sim}{x}\right) = c\int_{-\infty}^{\infty} \frac{1}{\left(\sigma^2\right)^{n/2}} \exp\left(\frac{-\beta}{2\sigma^2}\right) \exp\left\{-\frac{n}{2\sigma^2}\left(\mu - \frac{\sum_{i=1}^{n} \log x_i}{n}\right)^2\right\} d\mu$$

$$\pi\left(\sigma^2 \mid \underset{\sim}{x}\right) = \frac{c\exp\left(\frac{-\beta}{2\sigma^2}\right)}{\left(\sigma^2\right)^{n/2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{n}{2\sigma^2}\left(\mu - \frac{\sum_{i=1}^{n} \log x_i}{n}\right)^2\right\} d\mu$$

$$\pi\left(\sigma^2 \mid \underset{-}{x}\right) = c\frac{\exp\left(\frac{-\beta}{2\sigma^2}\right)\sqrt{2\pi}}{\left(\sigma^2\right)^{n/2}\sqrt{n}\sqrt{}}$$

$$\pi\left(\sigma^2 \mid \underset{-}{x}\right) = \frac{\exp\left(\frac{-\beta}{2\sigma^2}\right)\beta^{\frac{n-3}{2}}\sqrt{}}{\left(\sigma^2\right)^{n/2}2^{\frac{n-3}{2}}\Gamma\left(\frac{n-3}{2}\right)} \tag{14}$$

Posterior estimates of μ and σ^2

The marginal density of μ is given in (13) is a student's t pdf. Thus the posterior estimates of μ is given as

$$\mu^* = E\left(\mu \mid \underset{-}{x}\right) = \sqrt{\frac{n}{\beta}}\frac{1}{B\left(\frac{1}{2},\frac{n-3}{2}\right)}\int_{-\infty}^{\infty}\frac{\mu\, d\mu}{\left[1+\frac{n}{\beta}\left(\mu-\frac{\sum\log x_i}{n}\right)^2\right]^{\frac{n-2}{2}}}$$

Using the transformation $t = \sqrt{\frac{n}{\beta}}\left(\mu - \frac{\sum_{i=1}^{n}\log x_i}{n}\right)\sqrt{n-3}$

$$\mu^* = \frac{\sum_{i=1}^{n}\log x_i}{n\sqrt{n-3}\,B\left(\frac{1}{2},\frac{n-3}{2}\right)}\int_{-\infty}^{\infty}\frac{dt}{\left[1+\frac{t^2}{n-3}\right]^{\frac{n-2}{2}}}$$

$$\mu^* = \frac{\sum_{i=1}^{n} \log x_i}{n} \tag{15}$$

which is the posterior estimate for μ under uniform prior. Now the posterior estimate of σ^2 can be obtained from equation (14) as

$$\sigma^{*2} = \int_0^\infty \frac{\sigma^2 \exp\left(\frac{-\beta}{2\sigma^2}\right) \beta^{\frac{n-3}{2}}}{\left(\sigma^2\right)^{\frac{n-1}{2}} 2^{\frac{n-3}{2}} \Gamma\left(\frac{n-3}{2}\right)} d\sigma^2$$

$$\sigma^{*2} = \frac{\beta^{\frac{n-3}{2}}}{2^{\frac{n-3}{2}} \Gamma\left(\frac{n-3}{2}\right)} \int_0^\infty \frac{\exp\left(\frac{-\beta}{2\sigma^2}\right)}{\left(\sigma^2\right)^{\frac{n-3}{2}}} d\sigma^2$$

$$\sigma^{*2} = \frac{\beta}{n-5} \tag{16}$$

Thus, the posterior estimates of α_1 and β_1 are given by

$$\alpha_1^* = \exp\left[\mu^* + \frac{\sigma^{*2}}{2}\right] = \exp\left[\frac{\sum_{i=1}^{n} \log x_i}{n} + \frac{\beta}{2(n-5)}\right] \tag{17}$$

and

$$\beta_1^* = \exp\left[2\mu^* + \sigma^{*2}\right]\left[\exp\left(\sigma^{*2}\right) - 1\right]$$

$$\beta_1^* = \exp\left[2\frac{\sum_{i=1}^{n} \log x_i}{n} + \frac{\beta}{n-5} \right] \left[\exp\left(\frac{\beta}{n-5} \right) - 1 \right] \tag{18}$$

Simulation study and discussion

The estimates of the mean and variance using MLE and Bayesian estimation was obtained above. Next to obtain is the numerical relationship of point estimates using true value of the parameters, MLE and Bayesian estimation.

In this study, samples of 10, 20, 30, 40 and 50 observations were generated from lognormal pdf with parameters $\mu = 2$ and $\sigma = 1$. The simulations were done in R Software. The mean and variance were calculated to compare the methods of estimation. The results are presented in Table 1.

In Table 1, when point estimates of lognormal distribution are compared using true values of parameters with MLE and Bayesian estimation (by using uniform prior), the best estimator is the Maximum Likelihood (MLE) because it has the minimum variance.

Table 1. Point estimates of lognormal distribution compared using true values of parameters with MLE and Bayesian estimation

n	True values		MLE		Posterior estimates	
	Mean (α_1)	Variance (β_1)	Mean ($\hat{\alpha}_1$)	Variance ($\hat{\beta}_1$)	Mean (α_1^*)	Variance (β_1^*)
10			9.9004	21.336	9.1225	62.1358
20			12.6952	72.127	9.8447	130.857
30	12.1825	255.011	12.6655	52.1804	10.5913	70.6613
40			12.6452	56.9317	10.2974	71.6757
50			20.4039	267.461	12.5356	356.339

References

Aitchison, J., & Brown, J. A. C. (1957). *The lognormal distribution: With special reference to its uses in economics*. Cambridge: University Press.

Brownlee, K. A. (1949). *Industrial experimentation*. 2nd ed. Brooklyn: Chemical Pub. Co.

Delaporte, P. (1950). Etude statistique sur les proprietes des fontes. *Revue De L'institut International De Statistique / Review Of The International Statistical Institute*, (3/4), 161. doi:10.2307/1401035

Epstein. B. (1947). The mathematical description of certain breakage mechanism leading to the logarithmic-normal distribution. *Journal of Franklin Institute, 244*, 471-477.

Laplace, P. S. (1774). Mémorie sur la probabilitié des causes par les événemens. *Académie royale des sciences (France), 6*, 621-656.

Lindley, D. V. (1961). *Introduction to probability and statistics from a Bayesian viewpoint: Part 2, Inference*, Aberystwyth: University College of Wales.

Moroney, M. J. (1951). *Facts from figures*. Baltimore, MD: Penguin Books.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 314-335.

ISSN 1538 − 9472

Parameter Estimations Based On Kumaraswamy Progressive Type II Censored Data with Random Removals

Navid Feroze
Government Post Graduate College
Muzaffarabad
Azad Kashmir, Pakistan

Ibrahim El-Batal
Cairo University
Cairo, Egypt

The estimation of two parameters of the Kumaraswamy distribution is considered under Type II progressive censoring with random removals, where the number of units removed at each failure time has a binomial distribution. The *MLE* was used to obtain the estimators of the unknown parameters, and the asymptotic variance - covariance matrix was also obtained. The formula to compute the expected test time was derived. A numerical study was carried out for different combinations of model parameters. Different censoring schemes were used for the estimation, and performance of these schemes was compared.

Keywords: Expected test time, maximum likelihood estimation, progressive censoring, random removals

Introduction

Life tests are often one of the main research topics in many experimental designs. There are several situations in life testing, in reliability experiments and survival analysis in which units are lost or removed from the experiments while they are still alive. The loss may occur out of control or be reassigned. The out of control case can happen when an individual under study (testing) drops out. The other case may occur because of limitation of funds or to save time. For more details Balakrishnan and Aggarwala (2000) provide a comprehensive reference on the subject of progressive censoring and its applications. There are several types of censoring schemes; the Type II censoring scheme is one of the most common for consideration. In a Type II censoring, a total of *n* units are placed on test, but

Navid Feroze is a lecturer at the Government Post Graduate College Muzaffarabad, Azad Kashmir, Pakistan. Email him at: navidferoz@hotmail.com. Dr. El-Batal is a professor in the Institute of Statistical Studies and Research, Department of Mathematical Statistics. Email him at i_elbatal@yahoo.com.

instead of containing until all n units have failed, the test is terminated at time of the m_{th} ($1 \leq m \leq n$) unit failure. Type II censoring with different failure time distributions has been studied by many authors including Mann (1971), Meeker and Escobar (1991), and Lawless (2003).

If an experimenter desires to remove live units at points other than the n^{th} termination point of the life test, the above described scheme will not be of use to the experimenter. Type II censoring does not allow for units to be lost or removed from the test at other than the n_{th} termination point. This allowance may be desirable, as in the case of accidental breakage of experimental units, in which the loss of test units at points other than the termination point may be unavoidable. Intermediate removal may also be desirable when a compromise between reduced time of experimentation and the observation of at least some extreme lifetimes is sought. These reasons lead directly into the area of progressive censoring; see Balakrishnan and Agarwala (2000).

A generalization of Type II censoring is progressive Type II censoring. The progressive Type II censored life test is described as follows. Firstly, the experimenter places n units on a test at time zero, with m failures to be observed. When the first failure is observed, r_1 of the surviving units are randomly selected and removed. At the second observed failure, r_{th} of the surviving units are randomly selected and removed. This experiment terminates at the time when the m_{th} failure is observed and the remaining $r_m = n - r_1 - \ldots - r_{m-1} - m$ surviving units are all removed. The statistical inference on the parameters of failure time distributions under progressive Type II censoring has been studied by several authors, such as Cohen (1963), Mann (1971), Viveros and Balakrishnan (1974), Balakrishnan and Aggarwala (2000), Ng et al. (2002), Chan and Balakrishnan (2004), Soliman (2008) and Raqab et al. (2010) (and the references therein). Note that, in this scheme, r_1, r_2, ..., r_m are all pre-fixed. However, in some practical situations, these numbers may occur at random. Yuen and Tse (1996) indicated that, for example, the number of patients who drop out from a clinical test at each stage is random and cannot be pre-determined. In some reliability experiments, an experimenter may decide that it is inappropriate or too dangerous to carry on the testing on some of the tested units even though these units have not failed. In these cases, the pattern of removal at each failure is random. Suppose that any test unit being dropped out from the life test is independent of the others but with the same removal probability p. Then, Tse et al. (2004) indicated that the number of test units removed at each failure time has a binomial distribution. The main purpose of this article is to assess the required time to complete a life test under progressive Type II censored data with random removal (*PCR*). Assume that the

lifetime follows the Kumaraswamy distribution. The number of units removed at each failure time follows a binomial distribution with parameters *n* and *p*.

The Model

The maximum likelihood estimators for the parameters of the Kumaraswamy distribution are derived based on progressive Type II censoring. Let random variable *X* have a Kumaraswamy distribution with two positive shape parameters α and θ. The probability density function of *X* is given by

$$f(x,\alpha,\theta)=\alpha\theta x^{\alpha-1}\left(1-x^{\alpha}\right)^{\theta-1},0<x<1,\alpha,\theta>0 \tag{1}$$

while the cumulative distribution function is given by

$$F(x,\alpha,\theta)=1-\left(1-x^{\alpha}\right)^{\theta},0<x<1,\alpha,\theta>0 \tag{2}$$

Kumaraswamy (1980) was interested in distributions for hydrological random variables and actually proposed a mixture of a probability mass, at zero and density (1) over (0,1).

The corresponding survival function of random variable *X* is

$$\bar{F}(x,\alpha,\theta)=(1-x^{\alpha})^{\theta},0<x<1,\alpha,\theta>0 \tag{3}$$

and the failure (hazard) rate function takes the following form

$$h(x)=\frac{\alpha\theta x^{\alpha-1}(1-x^{\alpha})^{\theta-1}}{(1-x^{\alpha})^{\theta}}=\frac{\alpha\theta x^{\alpha-1}}{(1-x^{\alpha})} \tag{4}$$

For $X \geq 0$, let $X_1 < X_2 < ...< X_m$ be the *m* ordered failure times out of *n* randomly selected items, where *m* is predetermined before testing. At the i_{th} failure, r_i items are removed from the test. For progressive Type II censored sample with predetermined number of removals, say $R_1=r_1,R_2=r_2,.....,R_{m-1}=r_{m-1}$, where $R = R_1=r_1,R_2=r_2,.....,R_{m-1}=r_{m-1}$.

Let (X_1, X_2, ..., X_m) denote a progressive Type II censored sample. Then the joint probability density function of all m progressive Type II censored order statistic is given by

$$\begin{aligned} f_{X_1,X_2,...,X_m}(x_1,x_2,...,x_m) &= C\prod_{i=1}^{m} f(x_i)\left[1-F(x_i)\right]^{r_i} \\ &, x_1 < x_2 < ..., < x_m \\ \text{where} \quad C &= n\left(n-r_1-1\right)\left(n-r_1-r_2-1\right) \\ &...\left(n-r_1-r_2-r_3-...r_{m-1}-m+1\right) \end{aligned} \tag{5}$$

Thus for progressive Type II censoring with pre-determined number of removals $R = r$, the conditional likelihood function can be written as (Cohen, 1963)

$$\begin{aligned} L_1(x,\alpha,\theta|\mathrm{R}=\mathrm{r}) = C\prod_{i=1}^{m} f(x_i)\left[1-F(x_i)\right]^{r_i} = C\prod_{i=1}^{m}\left[\alpha\theta x_{(i)}^{\alpha-1}\left(1-x_{(i)}^{\alpha}\right)^{\theta-1}\right] \\ \left[\left(1-x_{(i)}^{\alpha}\right)\right]^{\theta r_i} \end{aligned} \tag{6}$$

Equation (6) is derived conditional on r_i, where r_i can be of any integer value between 0 and $n - m - (r_1 + r_2 + ...+ r_{m-1})$. The main difference between Type II progressive censoring and *PCR* is that the R are pre-determined in the former case while they are random in the latter case. Note that m is predetermined in both cases. Under *PCR*, the r_i terms are random. In particular, assume that each r_i follows a binomial distribution, such that

$$P\left(R_1 = r_1\right) = \binom{n-m}{r_1} p^{r_1}\left(1-p\right)^{n-m-r_1} \tag{7}$$

and

$$\begin{aligned} P\left(R_i = r_i, R_{i-1} = r_{i-1}, ..., R_1 = r_1\right) = \left(n-m-\sum_{j=1}^{i-1} r_j\right) \\ \times p^{r_i}\left(1-p\right)^{n-m-\sum_{j=1}^{i-1} r_j} \end{aligned} \tag{8}$$

where $0 \le r_i \le n-m-\left(r_1+r_2\ +...+r_{m-1}\right)$.

Furthermore, assume that R_i independent of X_i for all i. Then the likelihood function can be found as

$$L\left(x,\alpha,\theta,p \mid R=r\right)=L_1\left(x,\alpha,\theta \mid R=r\right)P\left(R,p\right) \tag{9}$$

where $P(R, p)$ is the is the probability distribution of the R terms ($R = r_1, r_2, \ldots, r_m$) and, in particular, results in

$$\begin{aligned} P\left(R,p\right) &= P\left(R_{m-1}=r_{m-1} \mid R_{m-2}=r_{m-2},...,R_1=r_1\right) \\ &\times P\left(R_{m-2}=r_{m-2} \mid R_{m-3}=r_{m-3},...,R_1=r_1\right) \\ &\times P\left(R_2=r_2 \mid R_{\ 1}=r_1\right)P\left(R_1=r_1\right) \end{aligned} \tag{10}$$

Substituting (4) and (5) into (7) results in

$$\begin{aligned} P\left(R,p\right) &= \frac{\left(n-m\right)!}{\Pi_{i=1}^{m-1}r_i!\left(n-m-\sum_{j=1}^{i-1}r_j\right)!}\,p^{\sum_{i=1}^{m-1}r_i} \\ &\times\left(1-p\right)^{(m-1)(n-m)-\sum_{i=1}^{m-1}(m-i)r_i} \end{aligned} \tag{11}$$

and

$$\begin{aligned} \log P\left(R,p\right) &= C+\sum_{i=1}^{m-1}r_i\log p \\ &+\left[(m-1)(n-m)-\sum_{i=1}^{m-1}(m-i)r_i\right]\log\left(1-p\right) \end{aligned} \tag{12}$$

Maximum Likelihood Estimation

The maximum likelihood estimators of the parameters α, θ, and p are derived based on progressive Type II censored data with binomial removals. Both point and interval estimations of the parameters are derived.

Point Estimations

Because $P(R, p)$ does not depend on the parameters α and θ, the maximum likelihood estimators (*MLE*s) of α and θ can be derived by maximizing (6) directly. Similarly, because $L_1(x, \alpha, \theta \mid R = r)$ does not involve the binomial parameter p, then the *MLE* of p can be found by maximizing $P(R, p)$ directly. The log likelihood function of (9) is given by

$$\log L\left(x,\alpha,\theta,p \mid R=r\right)=\log L_1\left(x,\alpha,\theta \mid R=r\right)+\log P\left(R,p\right)$$

where

$$\begin{aligned}\log L_1\left(x,\alpha,\theta \mid R=r\right)&=\log C+m\log\alpha+m\log\theta\\&\quad+\left(\alpha-1\right)\sum_{i=1}^{m}\log x_{(i)}\\&\quad+\left(\theta-1\right)\sum_{i=1}^{m}\log\left(1-x_{(i)}^{\alpha}\right)+\theta\sum_{i=1}^{m}r_i\log\left(1-x_{(i)}^{\alpha}\right)\end{aligned}\tag{13}$$

Take the partial derivative of $\log L_1(x, \alpha, \theta \mid R = r)$ with respect to α and θ and let them be zero

$$\begin{aligned}\frac{\partial\log\left(L_1\right)}{\partial\alpha}&=\frac{m}{\alpha}+m\log\theta+\sum_{i=1}^{m}\log x_{(i)}-\left(\theta-1\right)\sum_{i=1}^{m}\frac{x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)}\\&\quad-\theta\sum_{i=1}^{m}\frac{r_i x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)}=0\end{aligned}\tag{14}$$

$$\text{and }\frac{\partial\log(L_1)}{\partial\theta}=\frac{m}{\theta}+\sum_{i=1}^{m}\log\left(1-x_{(i)}^{\alpha}\right)+\sum_{i=1}^{m}r_i\log\left(1-x_{(i)}^{\alpha}\right)=0$$

$\widehat{\theta}=\dfrac{-m}{\sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right)}$ is given by $\widehat{\theta}$ of θ. Thus the *MLE* $\widehat{\alpha}$ of α and the *MLE* is the numerical solution of equation (14).

It is observed from (14) that the MLE of the parameter α cannot be obtained in closed form. It can be obtained by solving a one dimensional optimization problem. A simple fixed point iteration algorithm can be used to solve this

optimization problem. Firstly, the parameter θ in log-likelihood (13) has been replaced by its *MLE* the $\hat{\theta}$ resultant log-likelihood becomes

$$\log L_1(x,\alpha,\theta|\mathrm{R}=\mathrm{r}) = \log C + m\log\alpha + m\log\left\{\frac{-m}{\sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right)}\right\}$$
$$+(\alpha-1)\sum_{i=1}^{m}\log x_{(i)} + \left\{\frac{-m}{\sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right)}-1\right\}\sum_{i=1}^{m}\log\left(1-x_{(i)}^{\alpha}\right)$$
$$+\left\{\frac{-m}{\sum_{i=1}^{m}\left(r_i+1\right)\log(1-x^{\alpha}{}_{(i)})}\right\}\sum_{i=1}^{m}r_i\log\left(1-x_{(i)}^{\alpha}\right)$$

After some simplification it can be presented as

$$\log L_1(x,\alpha\mid R=r) \propto m\log\alpha + m\log\left\{\sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right)\right\}$$
$$+(\alpha-1)\sum_{i=1}^{m}\log x_{(i)} - \sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right) \qquad (15)$$

MLE of α can be obtained by maximizing (15) with respect to α and it is unique. Most of the standard iterative process can be used for finding the *MLE*. The following simple algorithm is proposed: If $\hat{\alpha}$ is the MLE of α, then it is obvious from $l'(\alpha) = \frac{\partial \log L_1(x,\alpha\mid R=r)}{\partial\alpha} = 0$ that $\hat{\alpha}$ satisfies the following fixed point type equation; $g(\alpha)=\alpha$

$$\frac{\partial \log L_1(x,\alpha \mid R=r)}{\partial \alpha} = \frac{m}{\alpha} - \frac{m\sum_{i=1}^{m}\frac{\left(r_i+1\right)x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)}}{\sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right)} + \sum_{i=1}^{m}\log x_{(i)} - \sum_{i=1}^{m}\frac{\left(r_i+1\right)x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)} = 0 \tag{16}$$

where

$$\alpha = g(\alpha) = \left[\frac{1}{m}\left\{\sum_{i=1}^{m}\frac{\left(r_i+1\right)x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)} - \sum_{i=1}^{m}\log x_{(i)}\right\} + \frac{\sum_{i=1}^{m}\frac{\left(r_i+1\right)x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)}}{\sum_{i=1}^{m}\left(r_i+1\right)\log\left(1-x_{(i)}^{\alpha}\right)}\right]^{-1}$$

The iterated result of the above function has been considered as an *MLE* of α and denoted by $\hat{\alpha}$. Now the approximate *MLE* of α has been incorporated in (14) to obtain the *MLE* of β.

Similarly, from (12) the partial derivative of log $P(R, p)$ with respect to binomial parameter p can be obtained by solving the following equation

$$\frac{\partial \log P(p,R)}{\partial p} = \frac{\sum_{i=1}^{m-1} r_i}{p} - \frac{(m-1)(n-m) - \sum_{i=1}^{m-1}(m-i)r_i}{1-p} = 0$$

thus the *MLE* of $\hat{p}$ of p is given by

$$\hat{p} = \frac{\sum_{i=1}^{m-1} r_i}{(m-1)(n-m) - \sum_{i=1}^{m-1}(m-i-1)r_i}$$

Interval Estimations

The approximate confidence intervals of the parameters based on the asymptotic distributions of the *MLE* of the parameters α, θ and p are derived in this subsection. The elements of the Fisher information matrix for the parameters of

the Kumaraswamy distribution based on progressive censored samples have been derived explicitly. The Fisher information matrix can be defined as

$$I(\alpha,\theta,p) = -\mathrm{E}\begin{bmatrix} \frac{\partial^2 \log(L)}{\partial \alpha^2} & \frac{\partial^2 \log(L)}{\partial \alpha\, \partial \theta} & 0 \\ \frac{\partial^2 \log(L)}{\partial \theta \partial \alpha} & \frac{\partial^2 \log(L)}{\partial \theta^2} & 0 \\ 0 & 0 & \frac{\partial^2 \log(L)}{\partial p^2} \end{bmatrix} = \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\theta} & 0 \\ I_{\theta\alpha} & I_{\theta\theta} & 0 \\ 0 & 0 & I_{pp} \end{bmatrix} \tag{17}$$

For the information matrix for α, θ and p, find

$$\frac{-\partial^2 \log(L)}{\partial \alpha^2} = \frac{m}{\alpha^2} + \sum_{i=1}^{m} \frac{x_{(i)}^{\alpha} \left(\log x_{(i)}\right)^2}{\left(1 - x_{(i)}^{\alpha}\right)^2} \left(\theta + \theta r_i - 1\right),$$

$$\frac{-\partial^2 \log(L)}{\partial \theta^2} = \frac{m}{\theta^2},$$

$$\frac{-\partial^2 \log(L)}{\partial \alpha\, \partial \theta} = \sum_{i=1}^{m} \frac{x_{(i)}^{\alpha} \log x_{(i)}}{\left(1 - x_{(i)}^{\alpha}\right)} + \sum_{i=1}^{m} \frac{r_i x_{(i)}^{\alpha} \log x_{(i)}}{\left(1 - x_{(i)}^{\alpha}\right)},$$

$$\frac{-\partial^2 \log(L)}{\partial P^2} = \frac{\sum_{I=}^{m-1} r_i}{p^2} + \frac{(m-1)(n-m) - \sum_{i=1}^{m-1} (m-i) r_i}{(1-p)^2},$$

and

$$\frac{-\partial^2 \log(L)}{\partial \alpha\, \partial p} = \frac{-\partial^2 \log(L)}{\partial \theta\, \partial p} = 0$$

In order to derive the expressions for $I_{\alpha\alpha}, I_{\alpha\theta}, I_{\theta\theta}$ **and** I_{pp} the distribution of the i^{th} order statistics from the Kumaraswamy distribution is required, which can be written as

$$g\left(x_{(i)}\right) = C_{n,i}\alpha\theta x_{(i)}^{\alpha-1}\left(1-x_{(i)}^{\alpha}\right)^{\theta-1} \times\left\{1-\left(1-x_{(i)}^{\alpha}\right)^{\theta}\right\}^{i-1}\left(1-x_{(i)}^{\alpha}\right)^{-\theta(n-i)},\ 0<x_{(i)}<1$$

where $C_{n,i} = \dfrac{n!}{(i-1)!(n-i)!}$

Here, the expectations necessary to derive the elements of the Fisher information matrix are

$$\mathrm{E}\left[\frac{x_{(i)}^{\alpha}\log x_{(i)}}{\left(1-x_{(i)}^{\alpha}\right)}\right] = C_{n,i}\alpha\theta\int_0^1 \log x_{(i)}\, x_{(i)}^{2\alpha-1}\left(1-x_{(i)}^{\alpha}\right)^{\theta(n+1-i)-2}\left\{1-\left(1-x_{(i)}^{\alpha}\right)^{\theta}\right\}^{i-1} dx_{(i)}$$

$$= C_{n,i}\alpha\theta\sum_{j=0}^{i-1}(-1)^j\binom{i-1}{j}\int_0^1 \log x_{(i)}\, x_{(i)}^{2\alpha-1}\left(1-x_{(i)}^{\alpha}\right)^{\theta(n+1-i+j)-2} dx_{(i)}$$

$$= C_{n,i}\alpha\theta\sum_{j=0}^{i-1}(-1)^j\binom{i-1}{j}\frac{1-HarmanicNumber\left[\theta(n+1-i+j)\right]}{4\alpha^2\theta(n+1-i+j)\left\{\theta(n+1-i+j)-1\right\}}$$

$$\mathrm{E}\left[\frac{x_{(i)}^{\alpha}\left(\log x_{(i)}\right)^2}{\left(1-x_{(i)}^{\alpha}\right)^2}\right] = C_{n,i}\alpha\theta\sum_{j=0}^{i-1}(-1)^j\binom{i-1}{j}\int_0^1\left(\log x_{(i)}\right)^2 x_{(i)}^{2\alpha-1}\left(1-x_{(i)}^{\alpha}\right)^{\theta(n+1-i+j)-3} dx_{(i)}$$

$$= C_{n,i}\alpha\theta\sum_{j=0}^{i-1}(-1)^j\binom{i-1}{j}\frac{\begin{array}{l}6(\gamma-2)\gamma \\ +\pi^2 \\ +6\psi\left(0,\theta(n+1-i+j)\right)\left\{2\gamma+\psi\left(0,\theta(n+1-i+j)\right)-2\right\}\end{array}}{48\alpha^3\left\{\theta(n+1-i+j)-1\right\}\left\{\theta(n+1-i+j)-2\right\}} - \frac{6\psi\left(1,\theta(n+1-i+j)\right)}{48\alpha^3\left\{\theta(n+1-i+j)-1\right\}\left\{\theta(n+1-i+j)-2\right\}}$$

where γ **and** $\psi(a,b)$ are Euler gamma and Poly gamma functions respectively. Using these results, the Fisher information matrix can be obtained, which can

further be used to derive the elements of the approximate variance-covariance matrix as

$$V=\begin{pmatrix} V_{\alpha\alpha} & V_{\alpha\theta} & 0 \\ V_{\theta\alpha} & V_{\theta\theta} & 0 \\ 0 & 0 & V_{pp} \end{pmatrix}=\begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\theta} & 0 \\ I_{\theta\alpha} & I_{\theta\theta} & 0 \\ 0 & 0 & I_{pp} \end{pmatrix}^{-1}$$

$\hat{\alpha},\hat{\theta},\hat{p}$. It is known that the asymptotic distribution of the *MLE* is given by

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\theta} \\ \hat{p} \end{pmatrix} \approx N\left[\begin{pmatrix} \alpha \\ \theta \\ p \end{pmatrix}, \begin{pmatrix} V_{\alpha\alpha} & V_{\alpha\theta} & 0 \\ V_{\theta\alpha} & V_{\theta\theta} & 0 \\ 0 & 0 & V_{pp} \end{pmatrix}\right] \tag{18}$$

Because V involves the parameters α, θ and p, replace the parameters by the corresponding maximum likelihood estimators in order to obtain an estimate of V, which is denoted by $\hat{V}$. Using (18), approximate 100(γ)% confidence intervals for α, θ and p are determined respectively as $\hat{\alpha}\pm Z_{\frac{\gamma}{2}}\sqrt{\hat{V}_{\alpha\alpha}}$, $\hat{\theta}\pm Z_{\frac{\gamma}{2}}\sqrt{\hat{V}_{\theta\theta}}$, $\hat{p}\pm Z_{\frac{\gamma}{2}}\sqrt{\hat{V}_{pp}}$, where z_γ is the upper 100(γ)% percentile of the standard normal distribution.

The Expected Time Test

In practical applications, it is often useful to have an idea of the test time of the whole test. For progressive Type II censoring sampling plan with random or binomial removals, the expected test time for the experiment is given by the expectation of the m^{th} order statistic $X_{(m)}$. From Balakrishnan and Aggarwala (2000), the conditional expectation of $X_{(m)}$ for a fixed set of $R = R_1 = r_1$, $R_2 = r_2$, $R_{m-1} = r_{m-1}$ is given by

$$E\left(X_{(m)} \mid R=r\right)=C(r)\sum_{l_1}^{r_1}\cdots\sum_{l_m}^{r_m}(-1)^{A}\frac{\binom{r_1}{l_1}\cdots\binom{r_m}{l_m}}{\prod_{i=1}^{m-1}h(l_i)}\left[\int_0^1 xf(x)F^{h(l_m)-1}(x)\,dx\right] \tag{19}$$

where

$$A=\sum_{i=1}^{m} l_i, C(r)=n(n-r_1-1)(n-r_1-r_2-2)\ldots\left(n-\sum_{i=1}^{m-1}(r_i+1)\right), h(l)=l_1+l_2+l_i+i$$

and '*i*' is the number of live units removed from experimentation (or number of failure units). Substituting (1) and (2) into (19) results in the following

$$E\left(X_{(m)} \mid R=r\right)=C(r)\sum_{l_1}^{r_1}\ldots\sum_{l_m}^{r_m}(-1)^A \frac{\binom{r_1}{l_1}\ldots\binom{r_m}{l_m}}{\prod_{i=1}^{m-1}h(l_i)}\left[\int_0^1 x\alpha\theta x^{\alpha-1}(1-x^\alpha)^{\theta-1}\right] \times\left[1-(1-x^\alpha)^\theta\right]^{h(l_m)-1}dx \quad (20)$$

Let

$$\begin{aligned}
S&=\left[\int_0^1 x\alpha\theta x^{\alpha-1}\left(1-x^\alpha\right)^{\theta-1}\right]\left[1-\left(1-x^\alpha\right)^\theta\right]^{h(l_m)-1}dx\\
&=\int_0^1 x\alpha\theta x^{\alpha-1}\left(1-x^\alpha\right)^{\theta-1}\sum_{k=0}^{h(l_m)-1}(-1)^k\binom{h(l_m)-1}{k}\left(1-x^\alpha\right)^{\theta k}dx\\
&=\theta\alpha\sum_{k=0}^{h(l_m)-1}(-1)^k\binom{h(l_m)-1}{k}\int_0^1 x^\alpha\left(1-x^\alpha\right)^{\theta k+\theta-1}dx\\
&=\theta\sum_{k=0}^{h(l_m)-1}(-1)^k\binom{h(l_m)-1}{k}\mathrm{B}\left(1+\frac{1}{\alpha},\theta\left(k+1\right)\right)
\end{aligned}$$

Plugging this quantity into the right hand side of equation (20), the expected test time of progressive Type II censoring with fixed number of removal will be

$$E(X_{(m)} \mid R=r)=\theta C(r)\sum_{l_1}^{r_1}\ldots\sum_{l_m}^{r_m}(-1)^A \frac{\binom{r_1}{l_1}\ldots\binom{r_m}{l_m}}{\prod_{i=1}^{m-1}h(l_i)} \times\sum_{k=0}^{h(l_m)-1}(-1)^k\binom{h(l_m)-1}{k}\mathrm{B}\left(1+\frac{1}{\alpha},\theta\left(k+1\right)\right) \quad (21)$$

Also, the expected time under the Type II censoring scheme without removal is defined by the expected value of the m_{th} failure time, denoted by $X^*(m)$ where

$$E\left(X^*(m)\right) = m\theta\binom{n}{m}\sum_{k=0}^{m-1}(-1)^k\binom{m-1}{k}\mathrm{B}\left(1+\frac{1}{\alpha},\theta(k+1)\right) \tag{22}$$

Because $r_i = 0$ for all $i = 1, 2, \ldots, m-1$. Similarly, the expected value of $X_{(m)}$ for complete sample can be derived from (22) by setting $m = n$ and $r_i = 0$ as

$$E\left(X^{**}(m)\right) = n\theta\sum_{k=0}^{m-1}(-1)^k\binom{n-1}{k}\mathrm{B}\left(1+\frac{1}{\alpha},\theta(k+1)\right) \tag{23}$$

Under *PCR*, the R terms are random. The expected time to complete an experiment under *PCR* is given by taking the expectation of both sides equation (21) with respect to the R terms. That is

$$E\left[(X_{(m)})\right] = E_R E\left[(X_{(m)}) \mid R = r\right] = \sum_{r_1=0}^{g(r_1)}\sum_{r_2=0}^{g(r_2)}\cdots\sum_{r_{m-1}=0}^{g(r_{m-1})} P(R)E\left[(X_{(m)}) \mid R = r\right] \tag{24}$$

where $g(r_i) = n - m - (r_1 + r_2 + \ldots + r_{i-1})$ and $P(R)$ is given in (10). Thus equation (24) gives an expression to compute the expected time for given values of m and n. To see how much time is saved under Type II progressive censoring, compare equations (23) and (24) where the ratio of the expected test time for Type I progressive censoring sample with binomial removals (*PCR*) with respect to the expected time for complete sample , that is

$$REET_1 = \frac{E\left[\left(X_{(m)}\right)\right] \text{ under PCR for a sample size } n}{E\left(X_{(m)}^{**}\right) \text{ under complete sampling for a sample size } n} \tag{25}$$

If replacing the numerator by the expected test time under Type II progressive censoring with random removals (*PCR*), this ratio is defined by $REET_2$. Notice that the ratios $REET_1$ and $REET_2$ provide important information in determining the shortest experiment time significantly if the sample size n is large. When $REET_1$ and $REET_2$ are closer to one, the test time under respective

censoring scheme is closer to the complete sample. The influence of the binomial probability removals p on the expected time can be studied by analyzing $REET_1$ for various values of p. The comparisons between the three expected times will be made in order to reward some information about m and n on the duration of the experiment. As it seems, analytical comparisons between these three expected times is difficult. Therefore, these comparisons can be made numerically for various values of m, n, α, and θ.

Numerical Study

The *MLE*s, their variances and 95% confidence intervals for parameters of the Kumaraswamy distribution using progressively censored data with random removals are now computed. The computations were made for different censoring schemes including various choices of m and n. the parametric space includes $(\alpha,\theta) \in \begin{Bmatrix} (0.50,0.75),(1.00,0.80),(2.00,3.00), \\ (2.50,1.50),(5.00,3.00),(4.00,2.50) \end{Bmatrix}$.

The censoring schemes are framed as follows:

Scheme 1:
$n = 20$, $m = 15$,
$r_1 = \ldots = r_{14} = 0$,
$r_{15} = 5$

Scheme 2:
$n = 20$, $m = 15$,
$r_1 = \ldots = r_7 = r_9 = \ldots = r_{15} = 0$,
$r_8 = 5$

Scheme 3:
$n = 20$, $m = 15$,
$r_2 = \ldots = r_{15} = 0$,
$r_1 = 5$

Scheme 4:
$n = 20$, $m = 18$,
$r_1 = \ldots = r_{17} = 0$,
$r_{18} = 2$

Scheme 5:
$n = 20$, $m = 18$,
$r_1 = \ldots = r_8 = r_{11} = \ldots = r_{18} = 0$,
$r_9 = r_{10} = 1$

Scheme 6:
$n = 20$, $m = 18$,
$r_2 = \ldots = r_{18} = 0$,
$r_1 = 2$

Scheme 7:
$n = 30$, $m = 20$,
$r_1 = \ldots = r_{19} = 0$,
$r_{20} = 10$

Scheme 8:
$n = 30$, $m = 20$,
$r_1 = \ldots = r_{10} = r_{13} = \ldots = r_{20} = 0$,
$r_{11} = r_{12} = 5$

Scheme 9:
$n = 30$, $m = 20$,
$r_2 = \ldots = r_{20} = 0$,
$r_1 = 10$

Scheme 10:
$n = 30$, $m = 25$,
$r_1 = \ldots = r_{19} = 0$,
$r_{20} = 5$

Scheme 11:
$n = 30$, $m = 25$,
$r_1 = \ldots = r_{10} = r_{13} = \ldots = r_{20} = 0$,
$r_{11} = 2$, $r_{12} = 3$

Scheme 12:
$n = 30$, $m = 25$,
$r_2 = \ldots = r_{20} = 0$,
$r_1 = 5$

Scheme 13:
$n = 40$, $m = 30$,
$r_1 = \ldots = r_{29} = 0$,
$r_{30} = 10$

Scheme 14:
$n = 40$, $m = 30$,
$r_1 = \ldots = r_{14} = r_{17} = \ldots = r_{30} = 0$,
$r_{15} = r_{16} = 5$

Scheme 15:
$n = 40$, $m = 30$,
$r_2 = \ldots = r_{30} = 0$,
$r_1 = 10$

Scheme 16:
$n = 40$, $m = 36$,
$r_1 = \ldots = r_{35} = 0$,
$r_{36} = 4$

Scheme 17:
$n = 40$, $m = 36$,
$r_1 = \ldots = r_{17} = r_{20} = \ldots = r_{36} = 0$,
$r_{18} = r_{19} = 2$

Scheme 18:
$n = 40$, $m = 36$,
$r_2 = \ldots = r_{36} = 0$,
$r_1 = 4$

The notations used in the tables are

$V(\hat{\alpha})$: Variance of the estimator

$LL(\hat{\alpha})$: Lower limit of the confidence interval

$UL(\hat{\alpha})$: Upper limit of the confidence interval

Table 1. *MLE*s, their variances and 95% confidence intervals for parameters using $\alpha = 0.50$, $\theta = 0.75$

Schemes	$\hat{\alpha}$	$\hat{\theta}$	$V(\hat{\alpha})$	$V(\hat{\theta})$	$LL(\hat{\alpha})$	$UL(\hat{\alpha})$	$LL(\hat{\theta})$	$UL(\hat{\theta})$
1	0.674492	0.941582	0.088685	0.059105	0.090802	1.258182	0.465076	1.418088
2	0.675560	0.955438	0.090411	0.060857	0.086219	1.264902	0.471920	1.438956
3	0.689900	0.988406	0.096476	0.065130	0.081111	1.298689	0.488204	1.488608
4	0.640946	0.917092	0.087820	0.046725	0.060112	1.221780	0.493417	1.340766
5	0.640955	0.926610	0.088027	0.047700	0.059434	1.222475	0.498538	1.354682
6	0.649568	0.934912	0.088397	0.048559	0.066829	1.232306	0.503005	1.366819
7	0.615257	0.909485	0.073949	0.041358	0.082264	1.148250	0.510886	1.308085
8	0.620614	0.912566	0.080909	0.041639	0.063103	1.178126	0.512616	1.312516
9	0.632944	0.915290	0.084524	0.041888	0.063112	1.202776	0.514146	1.316433
10	0.581573	0.836671	0.071226	0.028001	0.058485	1.104661	0.508696	1.164646
11	0.584405	0.854518	0.072029	0.029208	0.058375	1.110434	0.519547	1.189490
12	0.602121	0.896171	0.073575	0.032125	0.070475	1.133767	0.544872	1.247470
13	0.546138	0.809125	0.058654	0.021823	0.071455	1.020821	0.519583	1.098667
14	0.546197	0.816111	0.059933	0.022201	0.066365	1.026028	0.524069	1.108152
15	0.553892	0.823240	0.064746	0.022591	0.055164	1.052621	0.528647	1.117833
16	0.510985	0.774778	0.038303	0.016674	0.127392	0.894579	0.521684	1.027872
17	0.532987	0.781340	0.042096	0.016958	0.130846	0.935128	0.526102	1.036577
18	0.536114	0.796164	0.048681	0.017608	0.103665	0.968562	0.536084	1.056245

Table 2. *MLE*s, their variances and 95% confidence intervals for parameters using $\alpha = 1.00$, $\theta = 0.80$

Schemes	$\hat{\alpha}$	$\hat{\theta}$	$V(\hat{\alpha})$	$V(\hat{\theta})$	$LL(\hat{\alpha})$	$UL(\hat{\alpha})$	$LL(\hat{\theta})$	$UL(\hat{\theta})$
1	1.232107	1.026742	0.103913	0.070280	0.600290	1.863925	0.507139	1.546344
2	1.238510	1.026802	0.105153	0.070288	0.602936	1.874084	0.507168	1.546435
3	1.242170	1.027064	0.106155	0.070324	0.603573	1.880767	0.507298	1.546830
4	1.218273	0.985327	0.099675	0.053937	0.599476	1.837070	0.530129	1.440525
5	1.220254	0.987677	0.099778	0.054195	0.601136	1.839372	0.531393	1.443960
6	1.220848	1.008086	0.103164	0.056458	0.591313	1.850382	0.542374	1.473797
7	1.174262	0.958474	0.095742	0.045934	0.567794	1.780731	0.538404	1.378543
8	1.190808	0.961994	0.097742	0.046272	0.578039	1.803578	0.540381	1.383606
9	1.215313	0.969807	0.098676	0.047026	0.599623	1.831002	0.544770	1.394844
10	1.132875	0.937474	0.083493	0.035154	0.566531	1.699219	0.569984	1.304964
11	1.141405	0.941570	0.089372	0.035462	0.555460	1.727350	0.572475	1.310666
12	1.163471	0.942070	0.092902	0.035500	0.566065	1.760876	0.572779	1.311362
13	1.113024	0.878943	0.080758	0.025751	0.556034	1.670015	0.564417	1.193469
14	1.128151	0.928314	0.081152	0.028726	0.569802	1.686501	0.596121	1.260507
15	1.129390	0.936065	0.081241	0.029207	0.570737	1.688044	0.601098	1.271031
16	1.110531	0.853722	0.076051	0.020246	0.570015	1.651048	0.574840	1.132605
17	1.111389	0.856941	0.078882	0.020399	0.560905	1.661872	0.577007	1.136875
18	1.111504	0.865018	0.080522	0.020785	0.555328	1.667680	0.582445	1.147590

Table 3. *MLE*s, their variances and 95% confidence intervals for parameters using $\alpha = 2.00$, $\theta = 3.00$

Schemes	$\hat{\alpha}$	$\hat{\theta}$	$V(\hat{\alpha})$	$V(\hat{\theta})$	$LL(\hat{\alpha})$	$UL(\hat{\alpha})$	$LL(\hat{\theta})$	$UL(\hat{\theta})$
1	2.405773	3.331680	0.173545	0.740006	1.589262	3.222285	1.645617	5.017742
2	2.420848	3.341618	0.175239	0.744427	1.600362	3.241334	1.650526	5.032709
3	2.424628	3.349518	0.180851	0.747951	1.591108	3.258149	1.654428	5.044608
4	2.363294	3.323579	0.153386	0.613677	1.595669	3.130919	1.788164	4.858994
5	2.390266	3.328627	0.161481	0.615542	1.602645	3.177887	1.790880	4.866374
6	2.395203	3.331073	0.172947	0.616447	1.580101	3.210306	1.792196	4.869950
7	2.309179	3.258572	0.131193	0.530914	1.599256	3.019103	1.830440	4.686704
8	2.313648	3.285888	0.133527	0.539853	1.597437	3.029859	1.845784	4.725992
9	2.325209	3.321773	0.135464	0.551709	1.603822	3.046596	1.865942	4.777604
10	2.254230	3.191529	0.110980	0.407434	1.601281	2.907178	1.940450	4.442609
11	2.255714	3.220759	0.121270	0.414931	1.573166	2.938263	1.958221	4.483296
12	2.278655	3.255155	0.125297	0.423841	1.584868	2.972441	1.979134	4.531176
13	2.180235	3.120957	0.102417	0.324679	1.552983	2.807487	2.004137	4.237778
14	2.184516	3.135531	0.102438	0.327719	1.557201	2.811832	2.013496	4.257567
15	2.253760	3.140167	0.103276	0.328688	1.623882	2.883637	2.016473	4.263862
16	2.104538	3.112679	0.082326	0.269133	1.542166	2.666911	2.095871	4.129487
17	2.133270	3.117755	0.099195	0.270011	1.515963	2.750577	2.099288	4.136221
18	2.177042	3.117986	0.100672	0.270051	1.555156	2.798928	2.099444	4.136528

Table 4. *MLE*s, their variances and 95% confidence intervals for parameters using $\alpha = 2.50$, $\theta = 1.50$

Schemes	$\hat{\alpha}$	$\hat{\theta}$	$V(\hat{\alpha})$	$V(\hat{\theta})$	$LL(\hat{\alpha})$	$UL(\hat{\alpha})$	$LL(\hat{\theta})$	$UL(\hat{\theta})$
1	2.747904	1.730365	0.239122	0.199611	1.789462	3.706346	0.854679	2.606050
2	2.757010	1.732250	0.240154	0.200046	1.796503	3.717517	0.855610	2.608889
3	2.792415	1.733128	0.241268	0.200249	1.829681	3.755148	0.856044	2.610212
4	2.708039	1.681290	0.227382	0.157041	1.773420	3.642658	0.904574	2.458007
5	2.735921	1.713229	0.228650	0.163064	1.798701	3.673141	0.921757	2.504700
6	2.739513	1.728698	0.237295	0.166022	1.784740	3.694287	0.930080	2.527315
7	2.704854	1.647548	0.191089	0.135721	1.848065	3.561643	0.925478	2.369617
8	2.705733	1.665756	0.214797	0.138737	1.797348	3.614119	0.935706	2.395806
9	2.705802	1.678866	0.221835	0.140930	1.782654	3.628950	0.943071	2.414662
10	2.677496	1.612695	0.171758	0.104031	1.865201	3.489791	0.980518	2.244871
11	2.691698	1.616866	0.182896	0.104570	1.853477	3.529919	0.983055	2.250678
12	2.700314	1.624152	0.183897	0.105515	1.859804	3.540824	0.987484	2.260819
13	2.601492	1.542551	0.139221	0.079315	1.870170	3.332814	0.990556	2.094546
14	2.628847	1.554005	0.165617	0.080498	1.831203	3.426491	0.997912	2.110099
15	2.652432	1.581927	0.167527	0.083416	1.850203	3.454662	1.015842	2.148012
16	2.564242	1.518715	0.118527	0.064069	1.889457	3.239027	1.022602	2.014829
17	2.585654	1.532091	0.126751	0.065203	1.887853	3.283454	1.031608	2.032574
18	2.590843	1.538084	0.128757	0.065714	1.887542	3.294144	1.035643	2.040524

Table 5. *MLE*s, their variances and 95% confidence intervals for parameters using $\alpha = 3.00$, $\theta = 5.00$

Schemes	$\hat{\alpha}$	$\hat{\theta}$	$V(\hat{\alpha})$	$V(\hat{\theta})$	$LL(\hat{\alpha})$	$UL(\hat{\alpha})$	$LL(\hat{\theta})$	$UL(\hat{\theta})$
1	5.516209	3.339823	0.350879	0.743628	4.355202	6.677216	1.649639	5.030007
2	5.532574	3.379539	0.352004	0.761419	4.369708	6.695441	1.669256	5.089822
3	5.540173	3.381127	0.356985	0.762135	4.369107	6.711238	1.670041	5.092213
4	5.484345	3.318767	0.332414	0.611901	4.354300	6.614389	1.785575	4.851959
5	5.505484	3.325015	0.337346	0.614207	4.367087	6.643880	1.788936	4.861093
6	5.509684	3.333368	0.343636	0.617297	4.360723	6.658645	1.793431	4.873306
7	5.312807	3.283075	0.317312	0.538929	4.208731	6.416883	1.844204	4.721946
8	5.462259	3.311419	0.321464	0.548275	4.350982	6.573536	1.860126	4.762713
9	5.466648	3.313941	0.322346	0.549110	4.353848	6.579448	1.861542	4.766339
10	5.241934	3.232154	0.287640	0.417873	4.190746	6.293123	1.965150	4.499159
11	5.252454	3.250657	0.292829	0.422671	4.191826	6.313081	1.976400	4.524915
12	5.271842	3.252974	0.308544	0.423274	4.183126	6.360557	1.977808	4.528140
13	5.189160	3.208891	0.269755	0.343233	4.171176	6.207144	2.060604	4.357177
14	5.203086	3.220785	0.276195	0.345782	4.173022	6.233150	2.068242	4.373328
15	5.221027	3.226427	0.277078	0.346994	4.189318	6.252735	2.071865	4.380989
16	5.160790	3.144943	0.251366	0.274741	4.178117	6.143463	2.117595	4.172291
17	5.185568	3.147283	0.265617	0.275150	4.175422	6.195714	2.119170	4.175395
18	5.187746	3.196740	0.266302	0.283865	4.176298	6.199194	2.152472	4.241009

Table 6. *MLE*s, their variances and 95% confidence intervals for parameters using $\alpha = 4.00$, $\theta = 2.50$

Schemes	$\hat{\alpha}$	$\hat{\theta}$	$V(\hat{\alpha})$	$V(\hat{\theta})$	$LL(\hat{\alpha})$	$UL(\hat{\alpha})$	$LL(\hat{\theta})$	$UL(\hat{\theta})$
1	4.388710	2.724225	0.315652	0.494760	3.287524	5.489896	1.345577	4.102873
2	4.419342	2.761146	0.318780	0.508262	3.312714	5.525969	1.363813	4.158478
3	4.444269	2.786841	0.325157	0.517766	3.326626	5.561911	1.376505	4.197178
4	4.362571	2.681995	0.292876	0.399617	3.301859	5.423284	1.442977	3.921014
5	4.366521	2.690268	0.295321	0.402086	3.301390	5.431652	1.447428	3.933109
6	4.380645	2.721175	0.309534	0.411378	3.290184	5.471106	1.464056	3.978294
7	4.274244	2.667550	0.270601	0.355791	3.254665	5.293822	1.498445	3.836655
8	4.315312	2.668255	0.274710	0.355979	3.288022	5.342602	1.498840	3.837669
9	4.347334	2.671510	0.281012	0.356848	3.308326	5.386341	1.500669	3.842352
10	4.212643	2.589071	0.267819	0.268132	3.198320	5.226967	1.574155	3.603987
11	4.242698	2.638366	0.268059	0.278439	3.227919	5.257478	1.604127	3.672606
12	4.254647	2.644628	0.270255	0.279762	3.235719	5.273574	1.607934	3.681322
13	4.160297	2.574998	0.252870	0.221021	3.174688	5.145906	1.653547	3.496450
14	4.198490	2.579545	0.259322	0.221802	3.200386	5.196594	1.656467	3.502624
15	4.205975	2.580193	0.266779	0.221913	3.193622	5.218328	1.656883	3.503503
16	4.097057	2.552605	0.234649	0.180994	3.147622	5.046492	1.718754	3.386456
17	4.106227	2.554844	0.237447	0.181312	3.151148	5.061306	1.720262	3.389426
18	4.159811	2.558519	0.239472	0.181834	3.200668	5.118953	1.722736	3.394302

Tables 1-6 include the maximum likelihood estimates (*MLE*s), the variances of *MLE*s, and 95% confidence intervals for the parameters of the Kumaraswamy distribution under progressively Type II censored samples using different parametric values for various censoring schemes. It has been observed that by increasing the sample size (keeping censoring rate fixed), the estimated value of the parameter become closer to the true value, the variances of the *MLE*s decrease and widths of 95% confidence intervals tend to be lesser. This is an indication that the estimators are consistent in nature. It can further be assessed that the censoring schemes, concerned with survivals from the right, result in more precise results than their counterparts. As expected, the increase in true parametric values leads to the slower convergence of the estimates along with larger variances of the

estimates which lean to increase the widths of the confidence intervals. The increase in censoring rate, that is, the smaller values of '*m*' has the same natural consequences. However, these negative impacts can be protected by employing larger ($n > 30$) sample sizes.

Conclusion

This study addressed the problem of estimation of parameters of the Kumaraswamy distribution under progressive censoring based on random removals. The maximum likelihood estimation was used to serve the purpose. The findings of the study indicate that the proposed estimators are consistent in nature. It is interesting to note that the removal of items from the right leads to the most efficient results.

References

Balakrishnan, N., & Aggarwala, R. (2000). Progressive Censoring. *Theory, Methods and Applications*, BirkhÄauser, Boston.

Cohen, A. C. (1963). Progressively censored samples in life testing. *Technometrics*, *5*, 327-33

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, *46*, 79-88.

Lawless, J. F. (2003). Statistical models and methods for lifetime data. 2nd edition. *John Wiley and Sons, Hoboken*, 630.

Mann, N. R. (1971). Best linear invariant estimation forWeibull parameter under progressive censoring. *Technometrics*, *13*, 521-534

Meeker, W. Q., & Escobar, L. A. (1991). Statistical methods for reliability data. *New York: John Wiley & Sons*

Ng, H. K. T., Chan, P. S., & Balakrishnan, N. (2002). Estimation of parameters from progressively censored data using EM algorithm. *Computational Statistics and Data Analysis*, *39*, 371-386.

Ng, H. K. T., Chan, P. S., & Balakrishnan, N. (2004), Optimal progressive censoring plansfor the Weibull distribution. *Technometrics*, *46*(4), 470-481.

Raqab, M. Z., Asgharzadeh, A., & Valiollahi, R. (2010). Prediction for Pareto distribution based on progressively Type-II censored samples. *Comput. Statist. Data Anal*., *54*, 1732-1743.

Soliman, A. A. (2008). Estimation for Pareto model using general progressive censored data and asymmetric loss. *Communications in Statistics—Theory and Methods*, *37*, 1353-1370.

Tse, W., Cersosimo, M. G., & Gracies, J. M. (2004). Movements disorders and AIDs: a review. *Parkinsonisum and related disorders*, *10*, 323-334.

Viveros, R., & Balakrishnan, N. (1994). Interval estimation of life characteristics from progressively Censored data. *Technometric*, *36*, 84-91.

Yuen, H. K., & Tse, S. K. (1996). Parameters estimation for Weibull distributed lifetime under progressive censoring with random removals. *Journal of Statistical Computation and Simulation*, *55*, 57-71.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 336-343.

ISSN 1538 − 9472

Discriminating Between Generalized Exponential Distribution and Some Life Test Models Based on Population Quantiles

B. Srinivasa Rao
R.V.R & J.C College of Engineering
Guntur, India

R. R. L. Kantam
Acharya Nagarjuna University
Guntur, India

A test statistic based on population quantiles using sample order statistics is suggested. The quantiles of the test statistics are evaluated for generalized exponential distribution. Similar test statistic based on moments of sample order statistic is referred and the proposed test formula is compared with it. Between the pairs of the above models it is established that the test formula proposed by us is more effective and useful than the formula based on the moments of order statistics as developed by Sultan (2007).

Keywords: Population quantiles, generalized exponential distribution

Introduction

The three-parameter generalized exponential (GE) distribution has its probability density function (pdf) as

$$f(x)=\frac{\alpha}{\sigma}\left(1-e^{-\left(\frac{x-\mu}{\sigma}\right)}\right)^{\alpha-1}e^{-\left(\frac{x-\mu}{\sigma}\right)},\ x>\mu,\mu>0,\alpha>0,\sigma>0 \qquad (1)$$

Its cumulative distribution function (cdf) is given by

$$F(x)=\frac{\alpha}{\sigma}\left(1-e^{-\left(\frac{x-\mu}{\sigma}\right)}\right)^{\alpha},\ x>\mu,\mu>0,\alpha>0,\sigma>0 \qquad (2)$$

The two-parameter GE distribution has its pdf as

Dr. Rao is an associate professor of statistics in the Department of Mathematics and Statistics. Email her at: boyapatisrinu@yahoo.com. R. R. L. Kantam is in the Department of Statistics. Email at: kantam.rrl@gmail.com.

$$f(x)=\frac{\alpha}{\sigma}\left(1-e^{-\left(\frac{x}{\sigma}\right)}\right)^{\alpha-1}e^{-\left(\frac{x}{\sigma}\right)},x>0,\alpha>0,\sigma>0 \quad (3)$$

Its cdf is given by

$$F(x)=\frac{\alpha}{\sigma}\left(1-e^{-\left(\frac{x}{\sigma}\right)}\right)^{\alpha},x>0,\alpha>0,\sigma>0 \quad (4)$$

The GE distribution was introduced by Gupta and Kundu (1999). It is compared with the two-parameter Gamma and Weibull distributions in Gutpa and Kundu (2001a). Different models of estimations are discussed in Gutpa and Kundu (2001b). Raqab and Ahsanullah (2001) and Raqab (2002) studied the properties of order and record statistics from the two-parameter GE distribution respectively. Discriminating between gamma and GE distribution were studied by Gutpa and Kundu (2004). Discriminating between lognormal and GE distribution was given in Kundu et al (2005). The expected values of order statistics may not always be available in numerical form nor analytically simple beyond a given sample size. However if the distribution function is invertible analytically the population quantile for any ‘n’ can be easily obtained. Also moment of order statistics are conceptually similar to the population quantiles with an admissible measure of closeness. Therefore, quantiles are used to develop the test statistic and to distinguish the GE distribution from other well-known life testing models. The proposed work is similar to that of Sultan (2007) wherein moments of order statistics are used to develop the test statistic with GE distributions null population. The aim of this article is to explore the usefulness of analytical expressions of population quantiles of GE distribution. In section 2 the GE distribution and its quantiles are developed. In section 3 the goodness of fit tests of the two-parameter and three parameter GE distribution are developed. Section 4 deals with the power of the proposed test procedure in comparison with that of Sultan (2007) with the same alternative populations. In section 5 the performance of quantiles of GE distribution is tested, and Section 6 contains concluding remarks.

The GE distribution and its quantiles

The p^{th} quantile of population is defined as the solution of the equation $F(x)=p_i$ where F(x) is the cdf given in (1.4). This is also called the standard population

quantile. If $x_1, x_2, ...x_n$ is an ordered sample of size n and $p_i = \frac{i}{n+1}$ then the solution of $F(x) = p_i$ is defined as i[th] population quantile corresponding to its order statistic x_i. In the sample this is denoted by d_i i.e., $F(d_i) = \frac{i}{n+1}$. Expected value of its order statistic in the sample is denoted by μ_i, the theory of order statistics indicate that μ_i, d_i can be approximated by each other. If the distribution function of the population is in a closed form, d_i's sometimes can be obtained more easily than α's , moments of order statistics. This possibility is explored in developing the proposed test statistics of this article. Given the form of $F(x)$ and a natural number n, d_i's can be obtained by inverting the population distribution function. This was done for the GE distribution with the shape parameter $\alpha = 0.5, 2.0$ and sample size $n = 10, 20, 25$. To make use of them in the proposed test statistic, the details are given in the following sections.

Goodness-of-Fit Test using quantiles

Test for two-parameter case Let $x_1, x_2, ...x_n$ denote a sample from two-parametric GE distribution. The correlation type goodness of fit test procedure in this case using quantiles can be formed as follows:

$H_0 : F$ is correct, that is $x_1, x_2, ...x_n$ have $GE(0, \sigma, \alpha)$ given in (4) versus $H_1 : F$ is not correct, that is $x_1, x_2, ...x_n$ have another cdf and the test statistic used to run the test is given by

$$T_1 = \frac{\Sigma_{i=1}^{n} x_i d_i}{\sqrt{\Sigma_{i=1}^{n} x_i^2 \Sigma_{i=1}^{n} d_i^2}} \tag{5}$$

The test statistic T_1 is simulated through Monte-Carlo method based on 10,000 simulations. Table 1 represents the percentages points of T_1 for sample sizes $n = 10, 20, 25$ and $\alpha = 0.5, 2.0$. It can be seen from the Table 1, the percentage points of T_1 follow the naturally expected order.

Table 1. Percentage points of T_1

α	n	0.50%	1%	2%	2.50%	5%	10%	20%	30%	40%	50%
0.5	10	0.9947	0.9931	0.991	0.9901	0.9865	0.9808	0.9708	0.9611	0.9499	0.9374
	20	0.9921	0.9903	0.9879	0.9865	0.9818	0.9747	0.963	0.9506	0.9375	0.9221
	25	0.991	0.989	0.986	0.9847	0.9796	0.9717	0.959	0.9462	0.9334	0.9189
2	10	0.9954	0.9944	0.9933	0.9928	0.9908	0.9877	0.9829	0.9782	0.9734	0.9674
	20	0.9944	0.9932	0.9918	0.9912	0.9891	0.9863	0.9818	0.9773	0.9723	0.9665
	25	0.9936	0.9925	0.9911	0.9906	0.9886	0.9856	0.9812	0.977	0.9722	0.9669

Test for the three-parameter case Let $x_1, x_2, ...x_n$ denote a sample from three-parametric GE distribution and let $Z_i = X_i - X_1$ and $v_i = d_i - d_i, i = 1, 2, ...n-1$. The correlation type goodness of fit test in this case using quantiles can be formed as follows:

$H_0 : F$ is correct, that is $x_1, x_2, ...x_n$ have $GE(0, \sigma, \alpha)$ given in (2) versus $H_1 : F$ is not correct, that is $x_1, x_2, ...x_n$ have another cdf and the test statistic used to run the test is given by

$$T_2 = \frac{\Sigma_{i=1}^{n} z_i v_i}{\sqrt{\Sigma_{i=1}^{n} z_i^2 \Sigma_{i=1}^{n} v_i^2}} \tag{6}$$

The statistic T_2 is simulated through Monte-Carlo method based on 10,000 simulations. Table 2 represents the percentages points of T_2 for sample sizes $n = 10, 20, 25$ and $\alpha = 0.5, 2.0$. It can be seen from the Table 2 the percentage points of T_2 follow the naturally expected order.

Table 2. Percentage points of T_2

α	n	0.50%	1%	2%	2.50%	5%	10%	20%	30%	40%	50%
0.5	10	0.9951	0.9934	0.9914	0.9905	0.9869	0.9813	0.9712	0.9615	0.9505	0.9382
	20	0.9923	0.9904	0.998	0.9865	0.9818	0.9748	0.963	0.507	0.9376	0.9222
	25	0.9912	0.9891	0.9861	0.9847	0.9797	0.9718	0.959	0.9462	0.9334	0.919
2	10	0.9968	0.9958	0.9947	0.9943	0.9928	0.9902	0.9857	0.9812	0.9767	0.971
	20	0.9953	0.9942	0.9929	0.9924	0.9905	0.9879	0.9834	0.979	0.974	0.9682
	25	0.9945	0.9935	0.9922	0.9917	0.9898	0.987	0.9826	0.9785	0.9737	0.9683

Power of the test

The power of the test is calculated by replacing $GE(\mu,\sigma,\alpha)$ random variates generator in the simulation program with generators from the alternative distributions including: normal, lognormal, Cauchy, Weibull and gamma. Based on different sample sizes and 10,000 simulations, the power is calculated to be

$$Power = \frac{\#\text{ of rejections of } H_0}{10,000}$$

Where H_0 is rejected if $T_1(T_2)$ greater than or equal to the corresponding percentage points given in Table 1 (Table 2 and $T_1(T_2)$ is evaluated from the alternative distributions. Table 3 and 4 represent the power of the test for the two-parameter and three-parameter cases, respectively. The different alternative distributions considered are: (i) normal distribution N(μ, σ), (ii) lognormal Ln(μ, σ), (iii) Weibull distribution with location parameter μ, scale parameter σ and shape parameter α, W(μ, σ, α), (iv) gamma distribution with location parameter μ, scale parameter σ and shape parameter k G(μ, σ, k) and (v0 Cauchy distribution with location parameter μ and scale parameter σ C(μ, σ). Table 3 and 4 indicate that the correlation test has good power to reject sample from the chosen alternative distributions.

Table 3. Power of the test of the two-parameter case $(\sigma = 1)$

		N(0,1)		W(0,1,3)		G(0,1,7)	
α	n	5%	10%	5%	10%	5%	10%
0.5	10	0.9227	0.745	1	1	0.9991	0.9946
	20	0.9999	0.9985	1	1	0.9845	0.9306
	25	0.9989	0.9986	1	1	0.9959	0.9857
2	10	0.9971	0.9999	0.9991	0.9947	1	0.9996
	20	0.9997	0.9996	1	0.998	0.9972	0.9944
	25	0.9998	0.9997	1	1	0.9995	0.9986

Table 4. Power of the test of the two-parameter case $(\mu = 0, \sigma = 1)$

		LN(1,5)		W(0,1,6)		C(0,1)	
α	n	5%	10%	5%	10%	5%	10%
0.5	10	0.975	0.937	1	1	1	1
	20	0.975	0.943	1	1	1	1
	25	0.982	0.959	1	1	1	1
2	10	1	1	1	1	1	1
	20	1	1	1	1	1	1
	25	1	1	1	1	1	1

Tables similar to that of 3 and 4 are available in Sultan (2007), evaluated using the moments of order statistics. By comparison, notice that the coverage probability given in the tables are uniformly larger than what are given in Sultan (2007). Therefore, the test statistic proposed based on the population quantiles is more powerful than that based on the moments of order statistics. Moreover, for a distributional GE distribution moments of order statistics are not available completely beyond a given sample size whereas population quantiles are available for any sample size provided the mathematical form of the cdf is analytically invertible. Therefore it can be concluded that the proposed test statistic T is more powerful than that of Sultan (2007).

Numerical Examples

In order to show the performances of the test of GE distribution in both cases (two-parameter and three-parameter), four sets of order statistics each of size 25 were simulated, they are

1. Sample from GE(0,1,2): two-parameter case of the GE distribution with scale parameter is equal to 1 and shape parameter is equal to 2
2. Sample from GE(1,1,2): three-parameter case of GE distribution with location parameter is equal to 1, scale parameter is equal to 1 and shape parameter is equal to 2.
3. Sample from G(0,2,2): gamma distribution with location parameter is equal to 0, scale parameter is equal to 2 and shape parameter is equal to 2.

4. Sample from GE(2,2,2): gamma distribution with location parameter is equal to 2, scale parameter is equal to 2 and shape parameter is equal to 2.

The above four order statistics samples with the analogous quantiles of order statistics from GE(0,1,2) are used to run the test. The results of the test at 5% significance level and at α=2 (whether accept (A) or reject (R) H_0) are given for different values in the following table.

Table 5. Results at 5% significance, $\alpha = 2$

Decision			
GE(0,1,2)	**GE(1,1,2)**	**G(0,2,2)**	**G(2,2,2)**
A	A	R	R

Conclusions

This article proposed a test formula parallel to the one developed by Sultan (2007). It was found to be simple and can be used for any sample size. Moreover, it is more effective with respect to power evaluation and coverage probabilities.

References

Gupta, D. R., & Kundu, D. (1999). Generalized exponential distribution. *Australia and New Zealand Journal of Statistics, 41*(2), 173-188.

Gupta, D. R., & Kundu, D. (2001a). Exponentiated exponential family:an alternative to gamma and Weibull distributions. *Biometrica Journal, 43*, 117-130.

Gupta, D. R., & Kundu, D. (2001b). Generalized exponential distributions: different methods of estimation. *Journal of Statistical Computation and Simulation, 69*, 315-338.

Gupta, D. R., & Kundu, D. (2004). Discriminating between gamma and generalized exponential distributions. *Journal of Statistical Computation and Simulation, 74*(2),107-121.

Kundu, D., Gupta, D. R., & Manglic, A. (2005). Discriminating between the lognormal and generalized exponential distributions. *Journal of Statistical Planning and Inference, 127*, 213-227.

Raqab, M. Z.(2002). Inferences for generalized exponential distribution based on record statistics. *Journal of Statistical Planning and Inference, 104*, 339-350.

Raqab, M. Z., & Ahsanullah, M. (2001). Estimation of the location and scale parameters of generalized exponential distribution based on record statistics. *Journal of Statistical Computation and Simulation, 69*,109-124.

Sultan, K. S. (2007). Order Statistics from the Generalized Exponential Distribution and Applications. *Communications in Statistics-Theory and Methods, 36*(7), 1409-1418.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 344-pp.

ISSN 1538 − 9472

Akaike Information Criterion to Select the Parametric Detection Function for Kernel Estimator Using Line Transect Data

Omar Eidous
King Abdulaziz University
Jeddah, Saudi Arabia

Samar Al-Salman
Yarmouk University
Irbid, Jordan

Among different candidate parametric detection functions, it is suggested to use Akaike Information Criterion (*AIC*) to select the most appropriate one of them to fit line transect data. Four different detection functions are considered in this paper. Two of them are taken to satisfy the shoulder condition assumption and the other two estimators do not satisfy this condition. Once the appropriate detection function is determined, it also can be used to select the smoothing parameter of the nonparametric kernel estimator. For a wide range of target densities, a simulation results show the reasonable and good performances of the resulting estimators comparing with some existing estimator, particularly the usual kernel estimator when the half normal model is use as a reference to select the smoothing parameter.

Keywords: Line transect sampling, Akaike Information Criterion, kernel method, smoothing parameter

Introduction

Line transect sampling is one of the popular sampling method adopted by ecologists to estimate the population density D of specific objects in a given region. The estimation procedure can be performed by walking a distance L following a specific line transect, counts the number objects being investigated and records the perpendicular distance, X from the detected object to the center of the line transect. Let $g(x)$ be the detection function of observing an object at distance X, then X will tend to have a probability density function $f(x)$ of the same shape as $g(x)$ but scaled so that the area under $f(x)$ equals unity. Buckland et al. (2001) and Burnham et al. (1980) constitute the key references for this distance sampling procedure.

Dr. Eidous is in the Department of Statistics. Email him at: omarm@yu.edu.jo. Samar Al-Salman is a graduate of the Department of Statistics. Email him at alsalman85@yahoo.com.

The first logical assumption related the detection function $g(x)$ indicates that $g(x)$ is monotonically decreasing function in x. The second important assumption is that $g(0) = 1$, which indicates the objects located on the center of line will never be missed. In other words, this condition means that the probability of detected an object given that its perpendicular distance is zero equals one. In addition to the previous two assumptions, some authors (see Mack and Quang, 1998) stated that, in many practical situations the shape of the detection function of the data should have a shoulder at distance $x = 0$. If that is required then it can be translated mathematically as $g'(0) = 0$. The condition $g'(0) = 0$ is known in the literature as the shoulder condition assumption. However, Buckland et al. (2001) pointed out that the shoulder condition assumption may not be satisfied for some cases in practice, especially for the experiment with small objects or the experiment that performed with existing a fog or a tall grass etc. If $g(x)$ is monotonically decreasing and $g'(0) = 0$ then this ensures that $f(x)$ is in turn monotonically decreasing with $f'(0) = 0$.

Burnham and Anderson (1976) gave the fundamental relation for estimating the density of objects in a specific area, which can be expressed as $D = E(n) f(0) / 2L$, and the general estimate for D is given by $\hat{D} = n\hat{f}(0) / 2L$, where $E(n)$ is the expected value of the number of detected objects n, and $\hat{f}(0)$ is an approximate sample estimator of $f(0)$ based on the n observed perpendicular distances $x_1, x_2, \ldots, x_n$. Hence, the key aspect in line transects sampling can be reduced to be the modeling of $f(x)$ as well as the estimation of $f(0)$.

Let $X_1, X_2, \ldots, X_n$ be a random sample of n perpendicular distances from unknown *pdf* $f(x)$. A parametric approach would involve by assuming that $f(x)$ is a member of a family of proper *pdf* of a known functional form but depends on an unknown parameter θ, where θ may take a vector value and should be estimated by using the perpendicular distances. A variety of approaches to estimate θ will lead to $\hat{f}(0) = f(0, \hat{\theta})$. In contrast to the parametric method, the nonparametric kernel method requires no assumptions about the form of $f(x)$. This method allows the data at hand to talk about themselves.

Given that the line transect data are available and their true *pdf* is unknown, our first aim in this paper is to choose the most appropriate *pdf* for these data by considering four logical parametric models. The Akaiki Information Criterion (*AIC*) is suggested for use to select the best parametric model. The second aim is to use the *AIC* to determine the best parametric model that can be used as a reference to determine the smoothing parameter of the kernel estimator of $f(0)$.

Some Parametric Estimators

A number of parametric models have been proposed in the literature for $f(x)$. The negative exponential model and the half normal model are the most prominent models. Gates et al. (1968) suggested the negative exponential model with detection function,

$$g_1(x) = e^{-x/\alpha}, \quad x \geq 0$$

The corresponding *pdf* is,

$$f_1(x) = \frac{1}{\alpha} e^{-x/\alpha}, x \geq 0 \tag{1}$$

The maximum likelihood (ML) method indicates that the ML estimator of $f(0)$ is $\hat{f}_1(0) = 1/\bar{X}$, where $\bar{X}$ is the sample mean. The detection function $g_1(x)$ (or the *pdf* $f_1(x)$) do not satisfy the shoulder condition, which minimizes the importance of utilizing this model in line transect sampling. In contrast to the exponential model, the half normal model (Burnham et al., 1980) satisfies the shoulder condition assumption. The half normal detection function is given by

$$g_2(x) = e^{-x^2/2\sigma^2}$$

and the *pdf* is

$$f_2(x) = \frac{2}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \text{x} \geq 0. \tag{2}$$

The ML estimator of $f(0)$ is $\hat{f}_2(0) = \left(\frac{2}{\pi T}\right)^{1/2}$ under the half normal model, where $T = \sum_{i=1}^{n} x_i^2 / n$ is the ML estimator of σ^2. Ababneh and Eidous (2012) suggested the weighted exponential detection function with the form,

$$g_3(x) = e^{-\theta x}\left(2 - e^{-\theta x}\right),$$

and the corresponding *pdf* is

$$f_3(x)=\frac{2\theta}{3}e^{-\theta x}\left(2-e^{-\theta x}\right), x\geq 0, \theta>0 \tag{3}$$

The parameter required to estimate is

$$f_3(0)=\frac{2\theta}{3}$$

The expected value of *X* based on Model (3) is 7/(6θ), which gives $\hat{f}_3(0)=7/9\bar{X}$ as the moment estimator for $f_3(0)$. The moment estimator for $f_3(0)$ is given in a closed form, while the maximum likelihood estimator needs a numerical method to find it. It is worthwhile to note that the Model (3) satisfies the shoulder condition assumption. That is, $f_3'(0) = 0$. Finally, Burnham et al. (1980) suggested the Reversed Logistic detection function, which is given by

$$g_4(x)=\frac{3e^{-0x}}{1+2e^{-0x}},$$

and the corresponding *pdf* is given by

$$f_4(x)=\frac{2\theta}{3\ln(3)}\frac{3e^{-\theta x}}{1+2e^{-\theta x}}, x\geq 0 \tag{4}$$

It is easy to verify that Model (4) does not satisfy the shoulder condition assumption. Based on Model (4), the parameter that to estimate is $f_4(0)=\frac{2\theta}{3\ln(3)}$. If one decides to use the moment estimator for $f_4(0)$, then he obtains $\hat{f}_4(0)=\frac{2}{3\ln(3)}\frac{1.3078}{\bar{X}}=\frac{0.7936}{\bar{X}}$. Again the ML estimator of θ based on Model (4) does not exist in closed form and consequently it is not exist in closed form for $f_4(0)$. Therefore, a numerical method is required to find the corresponding ML estimator.

The Nonparametric Kernel Estimator

Let X_1, X_2, …, X_n be n perpendicular distances (assumed to be independent and identically distributed) from a continuous probability density function $f(x)$. Because the perpendicular distances are nonnegative, the usual kernel estimator of $f(x)$ (Silverman, 1986 and Chen, 1996) is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{x-x_i}{h}\right) + K\left(\frac{x+x_i}{h}\right)\right\}, x \geq 0 \tag{5}$$

where h is called the smoothing parameter (or bandwidth) and K is a symmetric kernel function assumed to satisfy the following conditions

$$\begin{gathered}\int_{-\infty}^{\infty} K(u)du = 1 \\ \int_{-\infty}^{\infty} uK(u)du = 0 \\ \int_{-\infty}^{\infty} u^2 K(u)du = c \neq 0 < \infty, \text{ where } c \text{ is a constant}\end{gathered} \tag{6}$$

The kernel estimator of $f(0)$ is obtained by taking $x = 0$ in Equation (5), which gives

$$\hat{f}(0) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{-x_i}{h}\right) + K\left(\frac{x_i}{h}\right)\right\}.$$

Since K is a symmetric function (i.e., $K(-x) = K(x)$), then

$$\hat{f}(0) = \frac{2}{nh}\sum_{i=1}^{n}K\left(\frac{x_i}{h}\right) \tag{7}$$

If $f(x)$ has a second continuous derivative at $x = 0$ then under the assumption that $h \to 0$ and $nh \to \infty$ when $n \to \infty$, the bias and variance of $\hat{f}(0)$ are (Chen, 1996)

$$bias\left(\hat{f}\left(0\right)\right)=2hf^{(1)}\left(0\right)R_1\left(K\right)+h^2f^{(2)}\left(0\right)R_2\left(K\right)+o\left(h^2\right) \quad (8)$$

and

$$\text{var}\left(\hat{f}\left(0\right)\right)=\frac{4f\left(0\right)M_2\left(K\right)}{nh}+o\left(nh\right)^{-1} \quad (9)$$

where $f^{(1)}\left(0\right)$ is the i^{th} derivative of $f(x)$ at $x = 0$, $R_z\left(K\right)=\int_0^\infty u^zK\left(u\right)du$ and $M_2\left(K\right)=\int_0^\infty K^2\left(u\right)du$. Hence, if $f'(0) = 0$, then the bias convergence rate is $O(h^2)$, if not (i.e., $f'(0) \neq 0$), the bias convergence rate is only $O(h)$, which is slower than $O(h^2)$ as $h\rightarrow 0$.

The estimator of D by using the kernel method is now obtained by substituting the estimator $\hat{f}\left(0\right)$ from (7) back into the formula of $\hat{D}$.

The Optimal Smoothing Parameter

There are many kernel functions that satisfy Condition (6). Wand and Jones (1995) pointed out that there is very little to choose between the various kernel functions on the basis of the mean square error of the estimator. In other words, given that the kernel function that satisfies (6) is selected, then the performance of the kernel estimator remains almost the same as any other kernel estimator when the kernel function is changed. However, it becomes very well know that the way to select the smoothing parameter h is very sensitive on the performance of the kernel estimator (see for example, Gerard and Schucany, 1999 and Eidous, 2005). The popular method that used to select h using line transect data is the reference method. This method can be used by adopting the half normal detection function as a reference. Gerard and Schucany (1999) pointed out that this technique is very acceptable in line transect sampling and there is no need to adopt the other computational methods such as least squares cross validation and likelihood cross validation methods.

As opposed to referring to only the half normal detection function to compute h, the other detection functions as stated in the section on Parametric Estimators are introduced as references to select h. This gives a choice to select

the most appropriate model to select the smoothing parameter and then, as expected, to improve the performances of the kernel estimator.

As stated, the smoothing parameter *h* has a strong effect on the accuracy of the kernel estimator (7) as illustrated by examining Formulas (8) and (9). As they demonstrated, the choice of a large value of *h* gives a large bias and small variance and vice versa. The logical method to determine *h* is to find its optimal value that minimizes the asymptotic mean square error (*MSE*) of the estimator $\hat{f}(0)$. The formula of the asymptotic *MSE* of $\hat{f}(0)$ (based on (8) and (9)) is given by

$$\begin{aligned} MSE\left(\hat{f}(0)\right) &= \left[bias\left(\hat{f}(0)\right)\right]^2 + \mathrm{var}\left(\hat{f}(0)\right) \\ &= h^4\left(f''(0)\right)^2 R^2(K) + \frac{4}{nh} f(0) M_2(K) \end{aligned} \tag{10}$$

Formula (10) is obtained by assuming that $f'(0) = 0$. By differentiating both sides of (10) with respect to *h* and equating the resulting equation with zero, the value of *h* that minimize the asymptotic *MSE* of $f(0)$ can be obtained. This value is known as the optimal smoothing parameter with respect to the asymptotic *MSE*, which is given by

$$h = \left\{ \frac{f(0) M_2(K)}{R_2^{\,2}(K)\left(f''(0)\right)^2} \right\}^{1/5} n^{-1/5} \tag{11}$$

The smoothing parameter *h* is now computed by assuming a reasonable form for $f(x)$. Gerard and Schucany (1999) compared among different methods to compute *h* in practice. They recommended to use the half-normal *pdf* as a reference, i.e., they assumed that $f(x) = f_2(x)$ (see Formula (2)), which gives $f(0) = \frac{2}{\sigma\sqrt{2\pi}}$ and $f''(0) = \frac{-2}{\sigma^3\sqrt{2\pi}}$, where σ is now estimated by its maximum likelihood estimator $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2}}{n}}$. Now, assume that the kernel function is the standard normal, i.e.

$K(t) = N(0,1)$, then $h = 0.933\hat{\sigma}n^{-1/5}$. By adopting the same technique, the formulas of h for the other densities can be derived, and are stated as follows:

- If $f(x) = f_1(x)$, then $h = 0.8918\hat{\theta}n^{-1/5}$, where $\hat{\theta} = \bar{X}$ is the maximum likelihood of θ.
- If $f(x) = f_3(x)$, then $h = 0.7330(1/\hat{\theta})n^{-1/5}$, where $\hat{\theta}$ is the maximum likelihood of θ under the weighted exponential *pdf*. However if the moments estimator of θ is required then $\hat{\theta} = \frac{7}{6\bar{X}}$.
- If $f(x) = f_4(x)$, then $h = 2.3734(1/\hat{\theta})n^{-1/5}$, where $\hat{\theta}$ is the maximum likelihood of θ under the reversed Logistic *pdf*. Note that the moments estimator of θ is $\hat{\theta} = \frac{1.3078}{\bar{X}}$.

Akaike Information Criterion (*AIC*) and the Proposed Estimators

The *AIC* (Buckland et al., 2001) is defined by

$$AIC = -2Log_e(L) + 2p$$

Where $\log_e(L)$ is the log-likelihood function evaluated at the maximum likelihood estimates of the model parameter and p is the number of parameters in the model. The above criterion provides a method to select the best model (among a set of models) that fit the data at hand. For a given data set, *AIC* is computed for each model and the model with the smallest *AIC* is considered to be better than the others. For models (1), (2), (3), and (4), the *AIC*s are given by

- $AIC_1 = 2n + 2nLog_e(\bar{X}) + 2$ for the negative exponential (Model 1).
- $AIC_2 = -2n\log_e(2) + n\log_e(2\pi\hat{\sigma}^2) + n + 2$ for the half normal (Model 2), where $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$ is the maximum likelihood estimator for σ^2.

- $AIC_3 = -2\left\{n\log_e\left(2\hat{\theta}/3\right) - \hat{\theta}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\log_e\left(2-\exp\left(-\hat{\theta}x_i\right)\right)\right\} + 2$ for the weighted exponential (Model 3), where $\hat{\theta}$ is the maximum likelihood estimator for θ.
- $AIC_4 = -2\left\{n\log_e\left(2\hat{\theta}\right) - n\log_e\left(\log_e(3)\right) - \hat{\theta}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\log_e\left(1+2\exp\left(-\hat{\theta}x_i\right)\right)\right\} + 2$ for the reversed Logistic (Model 4), where $\hat{\theta}$ is the maximum likelihood estimator for θ.

Two proposed estimators will be constructed for $f(0)$ based on the *AIC*. For a random sample of n perpendicular distances $X_1, X_2, \ldots, X_n$, the first proposed estimator is constructed by computing the *AIC* for each model and the model with the smallest *AIC* is selected to estimate $f(0)$. If the selected model is $f_j(x)$, $j = 1,2,3,4$ then $\hat{f}_j(0)$ is the estimator of $f(0)$. For example, if $f_1(x)$ is selected based on the *AIC* then $\hat{f}(0) = 1/\bar{X}$ is the estimator of $f(0)$. The first estimator of $f(0)$ will be denoted by $\hat{f}_P(0)$, where the sub *P* stands for "parametric." The second estimator is the usual kernel estimator (Estimator 7) but here the smoothing parameter of the kernel estimator is computed by using the reference model that is selected based on the *AIC*. In other words, compute the *AIC* for the previous four models and then select the model that has the smallest *AIC*, then based on the selected model, use the corresponding optimal formula to compute *h*. For example, if $f_1(x)$ is selected based on the *AIC* then $h = 0.8918\bar{X}n^{-1/5}$. This value is substituted in Estimator (7), which enables us to compute its final value for a given data set. The second estimator of $f(0)$ is denoted as $\hat{f}_N(0)$, where the sub *N* stands for "non-parametric."

Simulation Study and Results

In order to assess the performances of the proposed estimators $\hat{f}_P(0)$ and $\hat{f}_N(0)$ of $f(0)$, discussed in the previous section, a simulation study is performed. For the sake of comparison, the usual kernel estimator $\hat{f}(0)$ with smoothing parameter $h = 0.933\hat{\sigma}n^{-1/5}$ (Gerard and Schucany, 1999) is also considered. Four target families were considered in the simulation. These families were chosen using the criterion that they are representative of many different shapes that might occur in

the field. The target models – not necessary the same as the four models discussed in the Introduction – that used to simulate the perpendicular distances are

1) Exponential Power (EP) family (Pollack, 1978)

$$f(x)=\frac{1}{\Gamma(1+1/\beta)}e^{-x^{\beta}},\ x\geq 0,\ \beta>1$$

2) Hazard-Rate (HR) family (Hayes and Buckland, 1983)

$$f(x)=\frac{1}{\Gamma(1-1/\beta)}\left(1-e^{-x^{-\beta}}\right),\ x\geq 0,\ \beta>1$$

3) Beta (BE) family (Eberhardt, 1968)

$$f(x)=(w+\beta)(1-x/w)^{\beta},\ 0\leq x\leq w,\ \beta>0$$

4) General Reversed Logistic (GRL) family (Burnham et al., 1980)

$$f(x)=\frac{\beta b}{\ln(1+b)(1+be^{-\beta x})},\ x\geq 0,\ \beta,b>0$$

Two target models with two values for parameter β are selected from each of the above families. The selected model is truncated at a distance w. The selected values for β and for w for each model are as follows: (β, w) = (1,5), (2,3) for EP family; (β, w) = (1.5,20), (2,12) for HR family; (β, w) = (10,5), (20,9) for BE family; and (β, b, w) = (6,10,1), (8,30,1) for GRL family. These models cover a wide range for the detection functions of perpendicular distances, which vary near zero from spike to flat. It is worthwhile to mention here that the Reversed Logistic model (i.e. $f_4(x)$) is a special case of the above GRL with b = 2. The target GRL models that selected to simulate the data are taken for b = 10,30, which differ in their shape for $f_4(x)$. This choice is made to avoid our knowledge of the true detection function of the perpendicular distances.

It should be remarked that the EP model with β = 1, BE family and the GRL family do not satisfy the shoulder condition assumption. These choices were made in order to assess the robustness of the considered estimators with respect to the violation of the shoulder condition assumption. Note also that the other considered models satisfy the shoulder condition assumption.

For each model and for sample sizes n = 50,100,200, one thousand samples of perpendicular distances were randomly drawn. For each model and for each

sample size, Tables 1 – 4 demonstrate the simulated value of the relative bias (*RB*); $RB = \left\{ E\left(\hat{f}\left(0\right)\right) - f\left(0\right)\right\} / f\left(0\right)$ and the relative mean error (*RME*); $RME = \sqrt{MSE\left(\hat{f}\left(0\right)\right)} / f\left(0\right)$ for each considered estimator.

Table 1. *RB*, *RME*, and *EFF* of the different estimators when the perpendicular distances are simulated from EP detection function

		***n* = 50**			***n* = 100**			***n* = 200**		
Estimator	**Parameters**	***RB***	***RME***	***EFF***	***RB***	***RME***	***EFF***	***RB***	***RME***	***EFF***
$\hat{f}(0)$		-0.322	0.337	**1.000**	-0.288	0.298	**1.000**	-0.265	0.272	**1.000**
$\hat{f}_P(0)$	$\beta = 1$ $w = 5$	-0.043	0.209	**1.614**	-0.031	0.166	**1.798**	-0.024	0.133	**2.038**
$\hat{f}_N(0)$		-0.221	0.259	**1.302**	-0.194	0.224	**1.332**	-0.177	0.192	**1.417**
$\hat{f}(0)$		-0.083	0.156	**1.000**	-0.070	0.120	**1.000**	-0.052	0.097	**1.000**
$\hat{f}_P(0)$	$\beta = 2$ $w = 2.5$	0.043	0.172	**0.908**	0.008	0.080	**1.511**	0.006	0.055	**1.765**
$\hat{f}_N(0)$		-0.087	0.173	**0.906**	-0.082	0.124	**0.971**	-0.058	0.098	**0.994**

Table 2. *RB*, *RME*, and *EFF* of the different estimators when the perpendicular distances are simulated from HR detection function

		***n* = 50**			***n* = 100**			***n* = 200**		
Estimator	**Parameters**	***RB***	***RME***	***EFF***	***RB***	***RME***	***EFF***	***RB***	***RME***	***EFF***
$\hat{f}(0)$		-0.474	0.485	**1.000**	-0.439	0.444	**1.000**	-0.398	0.401	**1.000**
$\hat{f}_P(0)$	$\beta = 1.5$ $w = 20$	-0.255	0.297	**1.633**	-0.269	0.284	**1.563**	-0.277	0.285	**1.411**
$\hat{f}_N(0)$		-0.301	0.333	**1.457**	-0.268	0.284	**1.562**	-0.227	0.238	**1.689**
$\hat{f}(0)$		-0.266	0.290	**1.000**	-0.215	0.231	**1.000**	-0.171	0.183	**1.000**
$\hat{f}_P(0)$	$\beta = 2$ $w = 12$	0.050	0.188	**1.536**	0.069	0.146	**1.581**	0.068	0.113	**1.612**
$\hat{f}_N(0)$		-0.119	0.190	**1.522**	-0.080	0.131	**1.760**	-0.050	0.094	**1.944**

Table 3. *RB*, *RME*, and *EFF* of the different estimators when the perpendicular distances are simulated from BE detection function

		n = 50			*n* = 100			*n* = 200		
Estimator	Parameters	*RB*	*RME*	*EFF*	*RB*	*RME*	*EFF*	*RB*	*RME*	*EFF*
$\hat{f}(0)$	$\beta = 10$ $w = 5$	-0.299	0.316	**1.000**	-0.271	0.281	**1.000**	-0.244	0.252	**1.000**
$\hat{f}_P(0)$		-0.064	0.233	**1.354**	-0.039	0.194	**1.451**	-0.014	0.147	**1.716**
$\hat{f}_N(0)$		-0.217	0.264	**1.197**	-0.184	0.217	**1.298**	-0.147	0.166	**1.516**
$\hat{f}(0)$	$\beta = 20$ $w = 9$	-0.317	0.333	**1.000**	-0.285	0.296	**1.000**	-0.257	0.264	**1.000**
$\hat{f}_P(0)$		-0.068	0.220	**1.518**	-0.036	0.168	**1.756**	-0.020	0.133	**1.984**
$\hat{f}_N(0)$		-0.229	0.266	**1.253**	-0.198	0.224	**1.318**	-0.171	0.184	**1.431**

Table 4. *RB*, *RME*, and *EFF* of the different estimators when the perpendicular distances are simulated from GRL detection function

		n = 50			*n* = 100			*n* = 200		
Estimator	Parameters	*RB*	*RME*	*EFF*	*RB*	*RME*	*EFF*	*RB*	*RME*	*EFF*
$\hat{f}(0)$	$\beta = 6$ $b = 10$ $w = 1$	-0.092	0.166	**1.000**	-0.087	0.132	**1.000**	-0.070	0.107	**1.000**
$\hat{f}_P(0)$		0.028	0.172	**0.968**	-0.013	0.075	**1.758**	-0.009	0.050	**2.155**
$\hat{f}_N(0)$		-0.098	0.181	**0.922**	-0.097	0.135	**0.975**	-0.073	0.107	**1.001**
$\hat{f}(0)$	$\beta = 8$ $b = 30$ $w = 1$	-0.058	0.150	**1.000**	-0.040	0.115	**1.000**	-0.035	0.094	**1.000**
$\hat{f}_P(0)$		0.101	0.168	**0.891**	0.088	0.115	**1.001**	0.080	0.095	**0.996**
$\hat{f}_N(0)$		-0.063	0.157	**0.953**	-0.045	0.114	**1.015**	-0.035	0.094	**1.000**

For simple comparison, compute the efficiency (*EFF*) of the proposed estimators $\hat{f}_P(0)$ and $\hat{f}_N(0)$ with respect to the classic kernel estimator $\hat{f}(0)$, which is given by

$$EFF = \frac{MSE\left(\hat{f}(0)\right)}{MSE\left(\hat{f}_{P \text{ or } N}(0)\right)}$$

Depending on the simulation results of Tables 1 – 4, several conclusions can be drawn by inspecting the results in regard to *RB*, *RME*, and *EFF*

- The *RB*s that associated with the proposed estimators $\hat{f}_P(0)$ and $\hat{f}_N(0)$ are generally smaller in their magnitude than that associated with the classic kernel estimator $\hat{f}(0)$.
- The *RME*s for different estimators decrease when the sample size increases. This is a strong sign for the consistency of these estimators.
- The performance of the classical kernel estimator seems to be reasonable for EP model with $\beta = 2$ and for GRL model comparing to the proposed estimator $\hat{f}_N(0)$, at which the performances of the two estimators are similar. However, $\hat{f}_N(0)$ beats $\hat{f}(0)$ for the other cases.
- By comparing between the two proposed estimators $\hat{f}_N(0)$ and $\hat{f}_P(0)$, the performance of the former one seem to be surprisingly for most considered cases especially when the sample size increases.

Generally, Tables 1 – 4 demonstrate clearly that there is a significantly improvement by applying the estimator $\hat{f}_P(0)$ or even $\hat{f}_N(0)$ instead of the classic kernel estimator $\hat{f}(0)$.

References

Ababneh, F. & Eidous, O. (2012). A Weighted Exponential Detection Function Model for Line Transect Data. *Journal of Modern Applied Statistical Methods, 11*(1), 144-151.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2001). *Introduction to distance sampling*. Oxford: Oxford University Press.

Burnham, K. P., & Anderson, D. R. (1976). Mathematical models for nonparametric inferences from line transect data. *Biometrics*, *32*, 325-336.

Burnham, K. P., Anderson, D. R., & Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. *Wildlife Monograph* 72.

Chen, S. X. (1996). A kernel estimate for the density of a biological population by using line transect sampling. *Applied Statistics, 4*5, 135-150.

Eberhardt, L. L. (1968). A preliminary appraisal of line transects. *Journal of Wildlife Management, 32***,** 82-88.

Eidous, O. M. (2005). Bias correction for histogram estimator using line transect sampling. *Environmetrics, 16***,** 61-88.

Gates, C. E., Marshall, W. H., & Olson, D. P. (1968). Line transect method of estimating grouse population densities. *Biometrics, 24*, 135-145.

Gerard, P. D., & Schucany, W. R. (1999). Local Bandwidth Selection for Kernel Estimation of Population Densities with Line Transect sampling. *Biometrics, 55***,** 769-73.

Hayes, R.J., & Buckland, S.T. (1983). Radial distance models for line-transect method. *Biometrics, 39*, 29-42.

Mack, Y. P., & Quang, P. X. (1998). Kernel methods in line and point transect sampling. *Biometrics, 54***,** 609-619.

Pollock, K. H. (1978). A family of density estimators for line transect sampling. *Biometrics*, *34***,** 475-478.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Wand, M.P., & Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 358-370.

ISSN 1538 − 9472

Bayesian Joinpoint Regression Model for Childhood Brain Cancer Mortality

Ram C. Kafle
University of South Florida
Tampa, FL

Netra Khanal
The University of Tampa
Tampa, FL

Chris. P. Tsokos
University of South Florida
Tampa, FL

The Bayesian approach of joinpoint regression is widely used to analyze trends in cancer mortality, incidence and survival data. The Bayesian joinpoint regression model was used to study the childhood brain cancer mortality rate and its average percentage change (APC) per year. Annual observed mortality counts of children ages 0-19 from 1969-2009 obtained from Surveillance Epidemiology and End Results (SEER) database of National Cancer Institute (NCI) were analyzed. It was assumed that death counts are probabilistically characterized by the Poisson distribution and they were modeled using log link function. Results were compared with the mortality trend obtained using joinpoint software of NCI.

Keywords: Bayesian statistics, brain cancer, joinpoint regression, mortality, SEER.

Introduction

The social and economic burden due to cancer is growing and is the major public health problem in the United States. Brain cancer (brain tumor and other central nervous system (CNS) cancers) is one of the leading cancers ranking the second largest cause of childhood death due to cancers. Based on 1975-2007 incidence data reported by Kohler, et al. (2011), 65.2 percent of the children with brain tumors are diagnosed with malignant tumor whereas the percentage in adult is only 33.7. According to National Cancer Institute (NCI), leukemias and the cancers of the brain and nervous system in children account for more than half of the new cases. Brain tumors are the most common solid tumors and are the second most common type of pediatric cancer. The central brain tumor registry of the United States reports that approximately 4,300 children younger than age 20 are expected to be diagnosed with primary malignant and non-malignant brain

Ram C. Kafle is a PhD candidate in Statistics. Email him at: rckafle@mail.usf.edu. Dr. Netra Khanal is an Assistant Professor of Mathematics. Email him at: nkhanal@ut.edu. Dr. Chris P. Tsokos is a Distinguished University Professor in Mathematics and Statistics. Email him at: ctsokos@usf.edu.

cancer in 2013. According to Kleihues, et al. (1993), the histological appearances of childhood brain tumors differ significantly from that of adult and are classified into several large groups. The overall distribution of these tumors also differ significantly (Peterson, et al., 2006; Pollack, 1994; Pollack, 1999). Ullrich and Pomeroy (2003) reported that the Pilocytic astrocytoma is the main histologic types in children CNS tumors with relatively high frequency of occurrence. According to Ries et al. (2007), the overall incidence for childhood brain cancer rose from 1975 to 2004 with the greatest increase occurring from 1983 through 1986. But, it is found that the mortality rates are continuously decreasing, with relatively higher rate from 1969 to 1980 and slower rate from 1980 onwards. These previous works provide motivation to study the mortality trend in childhood brain cancer using a statistical model that is based on realistic assumptions.

The main objective of this study is to give the reliable estimates of the measure of cancer mortality trend that provide up-to-date information and recent changes in childhood brain cancer. The joinpoint regression model is preferable when analyzing the trend for several years as it enables the identification points in the trend where the significant changes occur (Kohler, et al., 2011). If it is assumed that the joinpoints are random variables that can occur at any locations within the data range, the log likelihood is not differentiable with respect to break points suggesting that the Bayesian method is a more realistic approach. The actual Bayesian Joinpoint Regression Model will be solely based on Bayesian model selection criteria with the smallest number of joinpoints that accurately describe the Annual Percentage Change (APC) in the trend of mortality rates. Having good estimates of the mortality rates will allow the detection of points in time where significant changes occur and provide the best possible predictions. More practically, it helps to monitor the progress being made in childhood brain cancer, and evaluate the effectiveness of current treatment methods with respect to the mortality rate.

The history of joinpoint is not very long. In 1992, Charlin et al. developed hierarchical Bayesian analysis of changepoint problem in which they used an iterative Monte Carlo method. Kim et al. (2000, 2004) proposed a nonparametric approach which is widely used for analyzing and predicting the mortality and incidence data. NCI still uses this methodology, among others to find the trends in mortality, incidence, and survival of cancers in the United States. Tiwari et al. (2005) first developed a Bayesian model selection method for joinpoint regression. They discussed two criteria to select the best model, one with smallest BIC and other related to the Bayes factor. All of the previous studies assumed that the

errors are IID normal which is not always relevant with the real application data such as mortality and incidence of a specific disease in a population. This normality assumption is relaxed by Ghosh et al. (2009) proposing a Bayesian approach on parametric and semi-parametric joinpoint regression model. They introduced a continuous prior for the joinpoints induced by the Dirichlet distribution. The generalized linear model with log link function in joinpoint regression model that evaluates and incorporates the uncertainty in both model selection and model parameters has been recently introduced and implemented by Martinez-Beneito et al. (2011).

Studied here is the mortality trend of childhood brain cancer data obtained from SEER database of NCI. The total annual observed mortality counts of children below 20 years of age from 1969-2009 is extracted. Being rare events, assume the mortality counts are probabilistically characterized by the Poisson probability distribution and model them using log link function. The Bayesian joinpoint regression model discussed previously was used to obtain the mortality trend assuming that the break points are continuous over time. The joinpoint regression model is also fitted using the joinpoint software of NCI for the same data and compare these two results. Observe that the model using Bayesian approach describes the data very well giving best possible short term predictions and performs a better improvement over the existing methods.

Joint Point Model

Let $Y_i, i = 1, 2, ..., n$ be the number of mortality counts during a period of time t_i in a population. Let there be k change points that describe the behavior of the data, then the mean of the above outcome process can be expressed as the following generalized linear model

$$g\left[E\left(Y_i \mid t_i\right)\right] = \alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \beta_j (t_i - \tau_j)^+, \quad (1)$$

where $\bar{t}$ is the mean of t_i, and τ_j is the change point in the model and g is monotonic and differentiable function, called the link function. The value of $(t_i - \tau_j)^+$ is $(t_i - \tau_j)$ if $(t_i - \tau_j)^+ > 0$ and 0 otherwise. For example, if there is no breakpoint in the model then

$$g\left[E\left(Y_i \mid t_i\right)\right] = \alpha + \beta_0(t_i - \bar{t});$$

and if there is one break point, the model becomes

$$g\left[E\left(Y_i \mid t_i\right)\right] = \alpha + \beta_0(t_i - \bar{t}) + \beta_1(t_i - \tau_1)^+.$$

The model with no breakpoint is named as M_0, one breakpoint as M_1 and so on. There will be M_{k+1} nested models over the model space in total depending upon the number of breakpoints.

In the proposed model given in (1), α, and β_0 represent the common parameters where as $\beta_j^{'}s$ are non-common parameters that are model-specific. β_0 together with $\beta_j^{'}s$ gives the slope for the different models with at least one change point. To give the same meaning across different models for all common parameters, Martinez-Beneito et al. (2011) proposed an alternative parametrization imposing different conditions. This work is motivated by their work and follows the same parametrization.

The purpose of this study is to fit the joinpoint regression model for the childhood brain and other CNS cancer mortality counts. This model is based on its probabilistic framework that provides a reliable estimates of annual mortality trend. Because the behavior of the mortality count data in the population is a rare event, characterized by Poisson distribution $(Y_i, Poi(\lambda_i, i = 1, 2, \cdots, n))$, it is modeled using natural log link function. Hence, the model in the equation (1) becomes

$$\log(\lambda_i) = \log(n_i) + \alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \delta_j \beta_j B_{\tau_j}(t_i) \qquad (2)$$

where n_i is the total number of population at time t_i, $B_{\tau_j}(t)$ is the piecewise linear function reparametrized as in Martinez-Beneito et al. (2011), called as break-point centered at τ_j, and $\delta_j, j = 1, 2, \cdots, k$ are binary indicator variables for the inclusion or exclusion of the change points in the model i.e.

$$\delta_j = \begin{cases} 1 & \text{for each breakpoint} \\ 0 & \text{otherwise} \end{cases}$$

The above equation (2) leads to the following estimated rate:

$$E(r_i) = exp(\alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \delta_j \beta_j B_{\tau_j}(t_i)). \quad (3)$$

The annual percentage change(APC) is used to characterized the trends or the change in rates over time. APC from i^{th} year to $(i+1)^{th}$ year is given as

$$APC_i = \frac{r_{i+1} - r_i}{r_i} \times 100.$$

Because the model can choose an infinite number of breakpoints, it is necessary to impose some restrictions on the position of the change points in the model. This is done by assigning minimum gap of two years between two joinpoints starting after the first years and ending before the last two years.

The aim is to find the trend that describes the behavior of the data well. This will be carried out by detecting the points and their locations where the significant changes occur within the data range. Finding such locations in this model selection problem is carried out by using Bayes Factor in which data updates the prior odds to yield posterior odds. Bayes Factor summarizes the relative support for one model versus another for all competing models by selecting a model with highest posterior probability. Therefore, the posterior probability of each model is calculated and the one with highest posterior probability is selected as the best model.

The specification of priors plays a major role in Bayesian model selection problem. In an objective Bayes solution to the model selection problem, the nature of the posterior distributions depends upon the selection of priors and is very sensitive if there are non-common parameters in the models as explained in Berger and Pericchi (2001), and Bayarri and García-Donato (2008). Furthermore, the choice of improper or vague priors would lead to arbitrary Bayes Factor and make the result computationally challenging (see Charlin et al., 1992; Martinez-Beneito et al., 2011). For the commmon parameters α, and β_0, choose flat priors i.e. $\pi(\alpha, \beta_0) \propto 1$. For non-common parameters, the divergence-based (DB) priors introduced in Bayarri, et al. (2008) as a generalization of the ideas discussed in Zellner and Siow (1980), Jeffreys (1961), and Zellner (1984) and implemented in Martinez-Beneito et al. (2011) is considered. The parameter space for τ is bounded, and hence the default prior $\pi(\tau) \propto 1$ was chosen. Based on the nature of

δ, it is reasonable to choose independent Bernoulli priors with a probability of success p with hyper priors for p being $Beta(\frac{1}{2}, \frac{k-1}{2})$ where k is the number of joinpoints in the model.

In Bayesian paradigm, finding a good candidate model from a set of nested models can be computationally intensive. The high dimensionality of the integrals makes the model selection procedure even more complex. In choosing priors, the distribution of the posterior probability is not analytically tractable, thus Gibbs variable selection approach in WinBUGS software is used to select the best model with significantly minimum number of joinpoints that describes the trend. The process is carried out in such a way that if one more joinpoint is added in the model, the model becomes insignificant.

Results

To apply the model discussed, annually observed mortality counts for childhood brain and other CNS cancers from the Surveillance Epidemiology and End Results (SEER) database of National Cancer Institute (NCI) from 1969-2009 were used. The data were extracted from publicly used database of the SEER program 7.1.0 with the adjustments of Katrina/Rita population.

The joinpoint model is fitted using WinBUGS software. The model is described by four unknown joinpoints ($k = 4$) to identify the time where changes in the slope of child brain cancer mortality trend occurs. Two parallel chains using different initial values are implemented. Each chain is run for 150,000 iterations giving 50,000 iterations as burn-in period. The posterior inferences is based on 100,000 iterations for each chain combining total of 200,000 iterations for each of the parameters. The posterior summaries for the parameters are given in Table 1. Out of competing five nested models, the model selection procedure selected the model with one joinpoint as given in Figure 1 (left). For the selected model with one joinpoint, the posterior distribution of each of the parameters is observed by monitoring the trace, iterations, Monte Carlo errors, standard deviations, and density curves. The trace for each of the parameters satisfy the convergence criteria. Also, the Monte Carlo errors are within 0.1% of the posterior standard deviations.

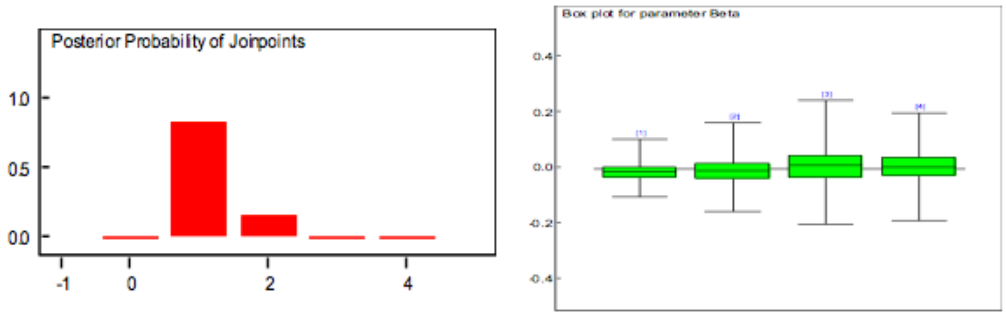


Figure 1: Posterior distribution of the number of joinpoints in child brain cancer mortality trend in United States (left), Box plot for parameters of joinpoints (right).

Table 1: Parameter estimates

node	mean	sd	MC error	2.50%	median	97.50%	start	sample
alpha	-11.76	0.006448	3.35E-05	-11.77	-11.76	-11.75	50000	200002
beta0	-0.01176	5.33E-04	2.79E-06	-0.01281	-0.01176	-0.01071	50000	200002
beta[1]	-0.0176	0.05287	7.68E-04	-0.09726	-0.02668	0.09301	50000	200002
beta[2]	-0.01679	0.09534	0.001723	-0.1736	-0.02925	0.1602	50000	200002
beta[3]	-0.00151	0.1265	0.001355	-0.218	-0.00167	0.2119	50000	200002
beta[4]	-7.90E-04	0.1114	0.001049	-0.1963	-1.52E-04	0.1938	50000	200002
delta[1]	0.5254	0.4994	0.01384	0	1	1	50000	200002
delta[2]	0.4684	0.499	0.01359	0	0	1	50000	200002
delta[3]	0.1156	0.3197	0.005156	0	0	1	50000	200002
delta[4]	0.05771	0.2332	0.001234	0	0	1	50000	200002

As depicted in the graph given in Figure 1 (left), the probability of the posterior distribution for one joinpoint is about 80%. The probability of the posterior distribution for no joinpoint is very low indicating that the linear trend is not a choice. Similary, the probability of posterior distribution does not support two, three, and four joinpoints as well. The boxplot for the parameters β_j, $j = 1, 2, 3, 4$ associated with change points is plotted in Figure 1 (right). Posterior means and 95% credible intervals of β_j's suggest that their posterior distributions are not discriminable. This indicates that no more than one joinpoint

is required and if more joinpoints are added, the model is not statistically significant.

The estimated rates for each year from 1969-2009 are obtained by averaging the estimates of joinpoint and other parameters in the model at every step of MCMC. The graph for the estimated rate and its prediction is given in Figure 2. The solid curve represents the estimated trend line for annual mortality rate whereas the dashed lines represent its 95% pointwise credible interval. The observed death rates are represented by unfilled circles. The extended graph beyond dashed vertical line represents the prediction of rate from 2010 to 2012. The predicted rates are obtained by averaging the joinpoint curve at every steps of the MCMC from the posterior predictive distribution.

The graph shows that the childhood cancer mortality rates declined faster from 1969 to 1978 compared to the rest of the time interval in a decreasing fashion. The overall mortality rate decreased from 1.056 to 0.63 per 100,000 by 2009 and is predicted to decrease continuously.

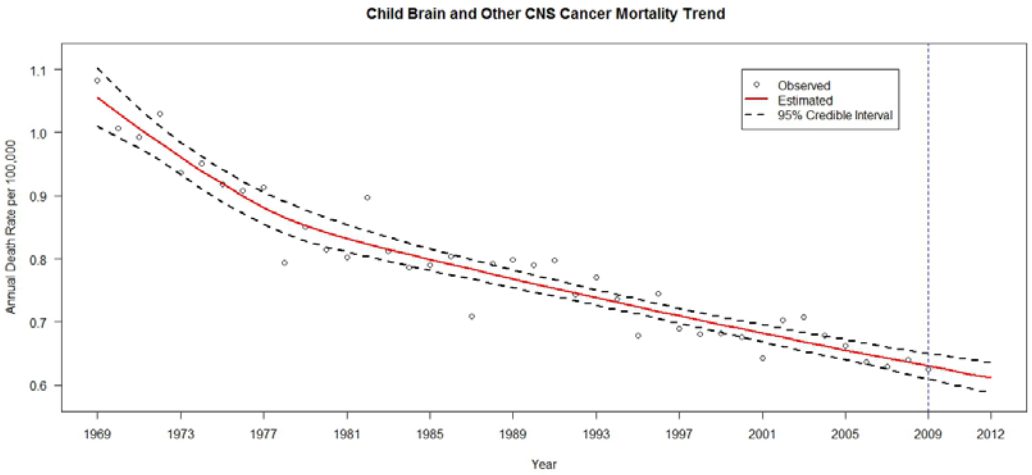


Figure 2: Estimated time trend for the annual observed mortality rate per 100,000 children

For the same data, the joinpoint regression model is fitted using the joinpoint software of NCI. The model was fitted with the assumption of Poisson variance using crude death rate with an autocorrelated errors based on the data. Here, the heteroscedasticity is conducted by joinpoint using weighted least square. Grid search method is used to select the joinpoint model with grid size of 2 years

leaving two years at the two ends of the data values. This was done to exactly match the condition imposed for identifiability problem in the Bayesian joinpoint model. The model selection method was performed using permutation test for four joinpoints which performs multiple tests to select the number of joinpoints using the Bonferroni correction at 0.05 overall significance level for multiple testing. The output is as shown in Figure 3.

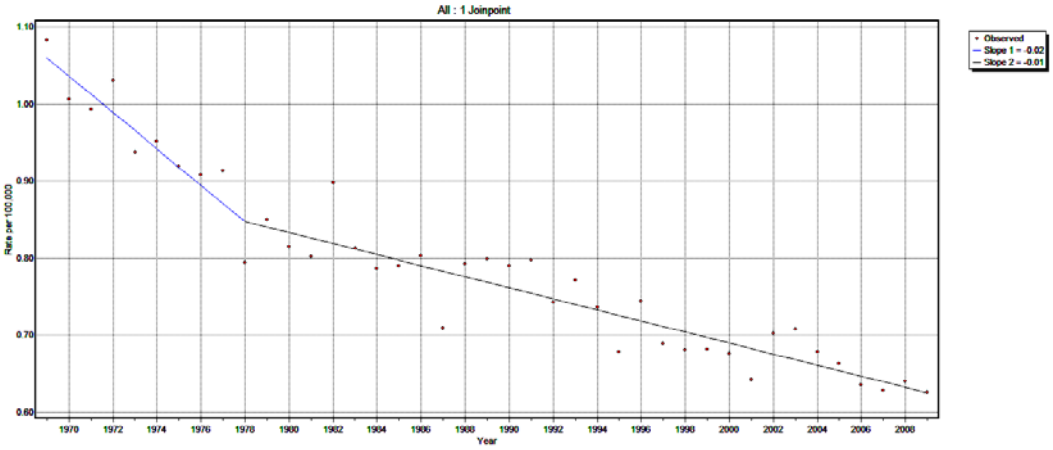


Figure 3: Mortality rates of child brain cancer(1969-2009) using the joinpoint software of NCI.

The solid line represents the estimated mortality rates obtained by using the joinpoint software of NCI. The graph shows that there is one joinpoint observed exactly at 1978. The trend line is piecewise linear indicating that the slopes of the rate curve before and after joinpoint are constant. It is not the case for the applied Bayesian joinpoint model as it gives the slope of the rate curve at any point. Also, the location of change point is discrete and occurs exactly at the whole number year in case of the regression trend given by joinpoint software whereas the location of change point is continuous in this case and can occur in between the years. Another difference is that the trend obtained from joinpoint software is descriptive but the regression trend obtained can give the insights for the mortality trend in future with credible bands.

The graph in Figure 4 gives the average rate of change in mortality rate per year from 1969 to 2009 and its predictions up to 2011. APC is approximately -2.31 for the first three years and increases from -2.29 in 1973 to -1.12 in 1980.

After 1980, APC looks almost contant with a fluctuation of 0.01 to 0.02 over the entire range. It means that the average rate of change per year in childhood brain cancer mortality rate has not been changed in recent years and is predicted to remain almost the same in the consequent years.

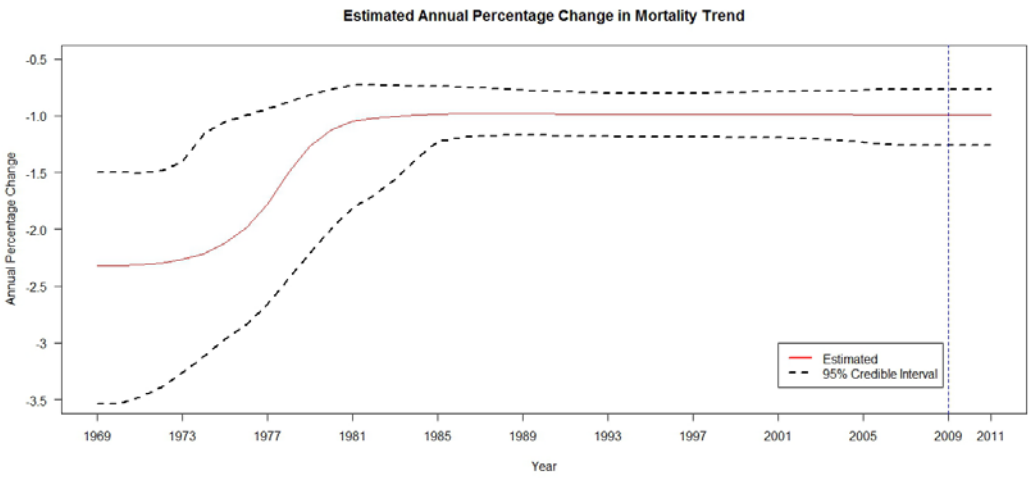


Figure 4: Estimated Annual Percentage Change in child brain cancer rates over time per 100,000 children

---

To check the validity, goodness of fit, and the assumptions of the proposed model, different model validation techniques discussed in literature are performed. The residual analysis is performed to check the robustness and fit of the developed model. The mean and standard deviation of the standardized residual are 0.000527 and 0.927 respectively. This indicates that the developed model fits the observed data well. The Chi-square statistics for the observed mortality data as well as for the predicated data in each iteration of MCMC are calculated.The difference between two statistics is monitored and their corresponding posterior p-value is obtained. The p-value based on the difference of Chi-squares obtained as a posterior mean using WinBUGS is 0.5513. The large p-value shows that the observed statistic is close from what is expected under the assumed model. Also, the observed mortality counts fall not only inside the 95% posterior intervals of replicated data but also close to their mean values indicating that the assumptions of Poisson distribution is valid.

Conclusion

This study applied newly developed Bayesian joinpoint regression model to uncover the patterns of childhood brain cancer mortality that provides an important information pertaining further study in the cases and control of the disease. Although, different studies have shown that the childhood cancer mortality rates continue to decline dramatically by more than 50% in the past two decades (Ries, et al., 2007; Kohler, et al., 2011) in the United States, only few studies have considered the probability distribution of the observed counts as Poisson and the location of the change points continuous in time. The application discussed here based on these probabilistic assumptions. The trend is obtained such that it describes the behavior of the observed data very well and gives the best possible short term predictions. The temporal trend provides the different slopes of the rate curve at each point of time. In contrast, the joinpoint software of NCI gives the same slope at each year between two change points. Also, it was possible to obtain the more accurate annual percentage change (APC) and it is observed that the APC is almost constant from 1981 and is predicted to remain constant. SEER routinely collects the data covering 28% of the US population and there is a three years lag in time to collect and process the data. In this scenario, predictions in the temporal trend and APC are very helpful to evaluate the effectiveness of the current status of the disease and play an important role to make evidence based policy. This improvement over the existing methods allow observation of the real progress being made in childhood brain cancer.

This work may be extended to study the influence in the mean of the outcome by incorporating covariates in the model. But the addition of covariates increases the complexity of the model. The Bayes Factors are sensitive to the prior specifications, and therefore further study is needed in selecting the objective priors by exploring different objective model selection criteria for priors that can deal with model uncertainty. Moreover, age standardized rates in this methodology could be a future extension. Also, studying incidence and mortality rates at the same time will depict the clear picture of real improvements being made in cancer research.

Acknowledgements

The authors wish to thank the University of Tampa Dana Foundation Grant.

References

Bayarri, M. J., & García-Donato, G. (2008). Generalization of Jeffreys' divergence based priors for Bayesian hypothesis testing. *Journal of the Royal Statistics Society; Series B (Statistical Methodology, 70*(5)**,** 981-1003.

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison (with discussion). *Model Selection, 38*, 135-207. Institute of Mathematical Statistics.

Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics, 41*(2), 389-405.

Ghosh, P., Basu, S., & Tiwari, R. C. (2009). Bayesian analysis of cancer rates from SEER program using parametric and semiparametric joinpoint regression models. *Journal of the American Statistical Association, 104*(486), 439-452.

Ghosh, K., & Tiwar, R. C. (2007). Prediction of U.S. cancer mortality counts using semiparametric Bayesian techniques. *Journal of the American Statistical Association, 102*(477), 7-15.

Jeffreys, H. (1961). Theorey of Probability. *London: Oxford University Press*, 3rd edition.

Kim, H., Fay, M. P., Feuer, E. J., & Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine, 19*(3), 335-351.

Kim, H. J., Fay, M., Yu, B., Barrett, M.J., & Feuer, E.J. (2004). Comparability of segmented line regression models. *Biometrics, 60*(4), 1005-1014.

Kleihues, P., Burgers, P. C., Scheithauer, B. W. , et al. (1993). *World health organization histological typing of tumors of the central nervous system*. New York: Springer-Verlag.

Kohler, B. A., Ward, E., McCarthy, B. J., et al. (2011). Annual report to the nation on the status of cancer, 1975-2007, featuring tumors of the brain and other nervous system. *Journal of National Cancer Institute, 103*(9), 714-736.

Levy, A. S. (2005). Brain tumors in children: evaluation and management. *Current Problems in Pediatric and Adolescent Health Care, 35*, 230-244.

Martinez-Beneito, M. A., Garcia-Donato, G., Salmeron, D. A. (2011). Bayesian joinpoint regression model with an unknown number of break-points. *Annals of Applied Statistics, 5*(3), 2150-2168.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New York: Wiley Publication.

Peterson, K. M., Shao, C., McCarter, R., MacDonald, T., & Byrne, J. (2006). An analysis of SEER data of increasing risk of secondary malignant neoplasmsamong long-term survivors of childhood brain tumors. *Pediatric Blood Cancer*, *47*(1), 83-88.

Pollack, I. F. (1994). Brain tumors in children. *The New England Journal of Medicine*, *331*(22), 1500-1507.

Pollack, I. F. (1999). Padiatric brain tumors. *Seminars in Surgical Oncology*, *16*(2), 73-90.

Ries, L., Melbert, D., & Krapcho, M. (2007). *SEER Cancer Statistics Review, 1975-2004*. National Cancer Institute.

Surveillance Epidemiology and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Mortality - All COD, Aggregated With State, Total U.S. (1969-2009) <Katrina/Rita Population Adjustment>.

Serveillance Research, National Cancer Institute, Joinpoint Regression program (serveillance.cancer.gov/joinpoint).

Tiwari, R. C., Cronin, K. C., Davis, W., Feuer, E. J., Yu, B., & Chib, S. (2005). Bayesian model selection for joinpoint regression with application to age-adjusted cancer rates. *Applied Statistics*, *54*(5), 919-939.

Ullrich, N. J., & Pomeroy, S. L. (2003). Pediatric brain tumors. *Neurologic Clinics*, *21*(4), 897-913.

Zellner, A. (1984). Posterior odds ratios for regression hypothesis: general considerations and some specific results. *Basic Issues in Econometrics*, 275-305. Chicago: University of Chicago Press.

Zelner, A., & Siow A. (1980). Posterior odds ratio for selected regression hypotheses. *In Bayesian Statistics 1* (J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith, eds.), *31*(1), 585-603. Valencia: University Press.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 371-380.

ISSN 1538 − 9472

Ordered Logit Regression Modeling of the Self-Rated Health in Hawai‘i, With Comparisons to the OLS Model

Hosik Min
University of South Alabama
Mobile, AL

Despite the ordinal nature of Self-Rated Health (SRH) variable, logistic regression models or regression models have been used without adequate justification for these applications. It is shown that ordered-logit regression model is the appropriate statistical strategy to estimate SRH, whereas the Ordinary LeastSquares model leads to biased conclusions.

Keywords: Ordered logit regression, OLS, ordinal outcome, self-rated health, health status

Introduction

Self-Rated Health (SRH) has long been a major research topic in health-related research (Mossey & Shapiro, 1982; Idler & Angel, 1990; Miilunpalo, Vuori, Oja, Pasanen, & Urponen, 1997; Eriksson, Unden, & Elofsson, 2001). The main reasons for this are that SRH can be used as an individual's general health status and/or an indicator of his or her quality of life and that the research importance of SRH will continue to increase because of a growing interest in health and healthy living (McMurdo, 2000; Eriksson, Unden, & Elofsson, 2001). Given the increased life expectancy and the aging of the population (NCHS, 2007), suffering and death from various diseases have declined, while the topic of healthy living has received greater attention (Row & Kahn, 1987; Glasgow, 2004; Glasgow, Min, & Brown, 2013). Health or lack thereof includes not only physical factors such as limitations to daily life activities (ADL) but also mental indicators such as SRH. As health condition and/or status can impact an individual's well-being in positive or negative ways, it is an important topic in public health.

Dr. Min is assistant professor in the Department of Sociology, Anthropology, and Social Work. Email him at: hksmin@gmail.com

Here the focus will be on methodological aspects; that is, the appropriateness of the ordered logit model for SRH, by comparing the results obtained using this method with those from the OLS model. SRH has often been measured as an ordinal variable; for instance, it is measured as a 5-point scale in this study (1=Poor, 2=Fair, 3=Good, 4=Very Good, and 5=Excellent). The analytical approach to handling this type of variable, however, is often logit regression (Avanath & Kleinbaum, 1997; Manor, Matthew, & Power, 2000; Pohlmann & Leitner, 2003) or Ordinary Least Squares (OLS) model (Winship & Mare, 1984; Wardle & Steptoe, 2003). The use of logit regression model can be easily denied because the logit model cannot deal with a dependent variable with more than two categorical and ordered outcomes in an appropriate way. In other words, if the SRH is developed as a dichotomous variable—e.g., poor versus good—and then a logit model is employed to estimate the logit coefficients, the results would lead to the loss of important information about the dependent variable (Hamilton, 1992; Berry, 1993; Hamilton, 1995; Avanath & Kleinbaum, 1997; Pohlmann & Leitner, 2003). In addition, only small percentage of Hawai‘i adults were having poor SRH (only 3%) in this study. Moreover, other kinds of social, cultural, and socioeconomic factors differentiating people who have good, very good, and excellent SRH will not be estimated if we use logit model.

Therefore, the goals of this paper are to present the methodological problems by comparing OLS, which often used to estimate ordinal outcome, and ordered logit models and to offer an easily understandable comparison of two methods by examining the likelihood of having a higher SRH in Hawai‘i. Considering wide use of OLS model for the dependent variable with many categories in ordered measurement (Mekelvey & Zavonia, 1975; Avanath & Kleinbaum, 1997), examining the statistical assumptions and violations the OLS model causes with ordered logit model would provide us a meaningful insights for employing an appropriate statistical methodology. In addition, this is a particularly important and relevant concern, given the expected increase in interest in general health status, both physical and mental.

As was indicated (Hawkes, 1971; Reynolds, 1973; Mekelvey & Zavonia, 1975; O’Brien, 1982), analyzing an ordinal variable with an ordinal regression model could lead to incorrect conclusions by violating the assumptions of the ordinal regression model. The OLS model has several assumptions known as a best linear unbiased estimating method (BLUE) (Hamilton, 1992; Berry, 1993; Hamilton, 1995; Avanath & Kleinbaum, 1997; Menard, 2001). For instance, the OLS model expects the dependent variable as linear and continuous one; the OLS model assumes that the mean of errors of prediction in the population regression

function must be zero; and the variance of the error term is constant for all values of independent variables, homoscedasticity

If the dependent variable is ordinal, however, these assumptions in general are not met (Mekelvey & Zavonia, 1975; Fox, 1991; Hamilton, 1992; Berry, 1993; Hamilton, 1995; Avanath & Kleinbaum, 1997). First of all, the ordinal dependent variable is non-linear, the values are presented in 0 to 1 probability as in a logit regression model; a non-linear model must have a different error structure and the error term does not have constant variance. As McKelvey and Zavoina (1975) argued, the OLS model may, in some cases, have the undesirable effect of causing regression analysis to severely underestimate the relative impact of certain variables. Accordingly, the ordered logit model, instead OLS model is considered to be the most appropriate methods if the dependent variable is ordinal to estimate more accurately (Hawkes, 1971; Reynolds, 1973; Mekelvey & Zavonia, 1975; O'Brien, 1982; Avanath & Kleinbaum, 1997; Pohlmann & Leitner, 2003).

Consequently, the best-fitting and most appropriate statistical model for handling the ordinal outcome is an ordered or probit model. This study, however, will use and focus on an ordered logit model, because the results of these two methods are similar and the ordered logit model is more common and its results are easier to interpret (Long & Freese, 2003).

Data and Methods

As described above, to measure the overall assessment of respondents' health, self-rated health (SRH) is used as a dependent variable. SRH is measured by a five-point scale and thus has a categorical and ordered nature. The best-fitting statistical model for handling the ordered outcome is known as an ordered-logistic regression model, which will be used as an analytical model here.

Here is an explanation of the ordered logit regression model. For the sake of explanation, symbols rather than actual variable names will be used (Long & Freese, 2003). Posit that Y is an ordinal dependent variable with c categories, and $\Pr(Y \le j)$ denotes the probability that the response on Y falls in category j or below (i.e., in category 1, 2, …, or j). This is called a cumulative probability. It equals the sum of the probabilities in category j and below:

$$\Pr(Y \le j) = \Pr(Y = 1) + (\Pr(Y = 2) + ... \Pr(Y = j) \quad (1)$$

A "c-category Y-dependent variable" has c cumulative probabilities: $\Pr(Y \le 1)$, $\Pr(Y \le 2)$, …, $\Pr(Y \le c)$. The final cumulative probability uses the entire scale; as a consequence, therefore, $\Pr(Y \le c) = 1$. The order of forming the final cumulative probabilities reflects the ordering of the dependent variable scale, and those probabilities themselves satisfy:

$$\Pr(Y \le 1) \le \Pr(Y \le 2) \le ... \le \Pr(Y \le c) = 1 \tag{2}$$

In an ordered logit model, an underlying probability score for an observation of being in the i^{th} response category is estimated as a linear function of the independent variables and a set of cut points. The probability of observing response category i corresponds to the probability that the estimated linear function, plus random error, is within the range of the cut points estimated for that response.

$$\begin{aligned} &\Pr\left(\text{Response Category for the j}^{th}\ \text{Outcome} = i\right) = \\ &\Pr\left(k_{i-1} < b_1 X_{1j} + b_2 X_{2j} + ... + b_k X_{kj} + u_j \le k_i\right) \end{aligned} \tag{3}$$

It is necessary to estimate the coefficients $b_1, b_2, ..., b_k$ along with cut points $k_1, k_2, ..., k_{i-1}$ where i is the number of possible response categories of the dependent variable. The coefficients and cut points are estimated using maximum likelihood.

To do this, the data used in this paper were obtained from the 2005 Hawaii Health Survey (HHS). The HHS is a representative-sample survey based on household, administered as a telephone interview survey to adult residents in more than 6,000 households each year. The principle objective of the survey is to provide statewide estimates of population parameters that describe (1) the current health status of the population; (2) respondents' access to and utilization of health care; and (3) the distribution of the population by age, sex, and ethnicity (SMS Research & Marketing Services, Inc., 2006).

The ordered logit regression model is thus estimated for the Hawai'i residents that predict their SRH using other socio-demographic and locale characteristics that have been shown in the demographic literature to be associated with SRH (Mossey & Shapiro, 1982; Idler & Angel, 1990; Kennedy, Kawachi, Glass, & Prothrow-Stith, 1998; Kawachi, Kennedy, & Glass, 1999;

Eriksson, Unden, & Elofsson, 2001). The controlling variables pertain to age, sex, race/ethnicity, marital status, education, and residential location. Some are measured as dummy variables and others as interval.

The variables are as follows: 1) Age is measured in years from age 18 to 99; 2) Male is a dummy variable indicating whether the respondent is male; if yes, it is coded as 1; 3) Married is a dummy variable indicating whether s/he is married; if yes, it is coded as 1; 4) Hawaiian is a dummy variable indicating whether the respondent is Native Hawaiian; if yes, it is coded as 1; 5) Japanese is a dummy variable indicating whether s/he is Japanese American; if yes, it is coded as 1; 6) Filipino is a dummy variable indicating whether the respondent is Filipino American; if yes, it is coded as 1; and 7) Other is a dummy variable indicating whether s/he belongs to Other ethnic categories; if yes, it is coded as 1 (with White used as the reference group); 8) Education is measured as 6 categories from illiterate to 4 or more years of college education (1=Illiterate/Only Kindergarten; 2=Grade 1 to 8; 3=Grade 9-11; 4=Grade 12 or GED; 5=College, 1 to 3 years; 6=College, 4 years or more); 9) Big Island is a dummy variable indicating whether the respondent lives in Big Island; if yes, it is coded as 1; 10) Kauaʻi is a dummy variable indicating whether s/he lives in Kauaʻi; if yes, it is coded as 1; 11) Maui is a dummy variable indicating whether the respondent lives in Maui; if yes, it is coded as 1 (with Oʻahu used as reference variable).

Results of Ordered Logit Regression Versus OLS Analysis

Table 1 presents frequency distributions for all independent variables as well as the dependent one. The average score of SRH for Hawaiʻi residents was 3.57, which lies between good and very good. The average age was 47.6 years old among the adult population (age 18 and over). Half of them were male (49%). Six out of ten Hawai'i adults were married (60%). As for race/ethnicity, 21% were Native Hawaiian, 22% were Japanese American, 15% were Filipino, and 17% were Other. The average level of education was 4.86, or close to 1-3 years of college education. As for residence, 13% lived in Big Island, 5% lived in Kauaʻi, 12% lived in Maui, and the remaining 70% lived in Oʻahu.

Table 2 presents the results of the ordered-logistic regression and the OLS analysis for Hawaiʻi adults in 2005. The results show that overall model fit was significant for both models, and most coefficients in both models were significant. The older the respondent, the lower the SRH; if s/he was married, s/he was more likely to have a higher SRH; compared to white respondents, all other racial and

ethnic categories, such as Native Hawaiian, Japanese American, Filipino American, and Other, show a lower likelihood of having a higher SRH. Also, as expected, the more educated the respondent, the higher the SRH; a person living in Kauaʻi and Maui has a higher likelihood of having higher SRH compared to a person living in Oʻahu.

Table 1. Descriptive Statistics from the 2005 Hawaii Health Survey (n=898, 593, weighted)

Variable	Mean	Std. Dev.
Self-rated Health	3.57	1.04
Age	47.60	17.59
Male	0.49	0.50
Marital Status		
Married	0.60	0.49
Race/Ethnicity		
Hawaiian	0.21	0.41
Filipino	0.15	0.35
Japanese	0.22	0.42
Other	0.17	0.38
Socioeconomic Status		
Education	4.86	1.02
Residence Island		
Big Island	0.13	0.34
Kauaʻi	0.05	0.22
Maui	0.12	0.32

The results, however, indeed present the evidence of inappropriateness of using OLS model compared to the ordered logit model. The male variable provided important information regardless of whether an ordered logit model or OLS was used to deal with an ordinal dependent variable. A male was shown to have a higher likelihood of having a higher SRH compared to female counterparts in the ordered logit regression model, but not in the OLS. As previous studies have pointed out, using an OLS model for an ordinal-dependent variable indeed produces this inconsistent and biased result: It could be concluded that male did

not have any effect on SRH, which would be crucially misleading in the OLS model. In addition, all the values of the coefficients in the OLS model were severely underestimated compared to those of the ordered logit model, which lessened the effects of contributing factors on SRH.

Table 2. Comparison of the Analysis Results of Ordered Logit Regression and OLS from 2005 Hawaii Health Survey (n=898, 593, weighted)

	Ordered Logit Regression			OLS		
Variable	**b**	**z**		**b**	**t**	
Age	-0.026	-220.65	*	-0.014	-232.6	*
Male	0.013	3.38	**	-0.003	1.68	
Marital Status						
Married	0.143	35.35	*	0.083	38.11	*
Race/Ethnicity						
Hawaiian	-0.583	-98.76	*	-0.298	-94.64	*
Japanese	-0.59	-10.18	*	-0.297	-97.14	*
Filipino	-0.642	-98.69	*	-0.315	-90.29	*
Other	-0.406	-65.85	*	-0.207	-62.84	*
Socioeconomic Status						
Education		162.95	*	0.176	165.57	*
Residence Island						
Big Island		0.71		0.005	1.52	
Kaua‘i		3.71	*	0.032	6.74	*
Maui		11.52	*	0.031	9.38	*
	LR Chi^2	106,789.24		F	10,379.23	
	Pseudo-R^2	0.043	*	Adj. R^2	0.113	*

* p<.05; ** p<.001
Note: The values of cut points to ordered logit regression and the value of constants for OLS are not shown here.

Discussion

This paper deals with an appropriate use of statistical modeling that frequently occurs when modeling ordinal variables, Self-Rated Health, which is measured using a 5-point scale here. By comparing the results of ordered logit regression and OLS models, this study could illustrate the potential problems with using OLS in the analysis of ordinal SRH variables. While most of the conclusions from the OLS model were similar to those from the ordered logit regression model, significant differences do exist. Most of all, the insignificance of male in the OLS model could lead to incorrect conclusions regarding this variable. In fact, the significant and positive effect for male had on a respondent's SRH score was revealed when this study used the ordered logit model. Furthermore, the OLS model underestimated the effects of all coefficients.

Accordingly, this study appears to show that the use of an ordered logit regression model is statistically appropriate for the modeling of Self-Rated Heath, which has an ordinal characteristic, in Hawai'i's adult population. More specifically, the use of the ordered logit regression model could help avoid inconsistent and biased conclusions and their detrimental effects on public health policy.

Considering the fact that the importance of studying health status indicators such as SRH continues to rise, the use of an appropriate analytical strategy will be invaluable in the future.

References

Avanath, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology, 26,* 1323-1333.

Berry, W. (1993). *Understanding regression assumptions*. 1st ed. Sage Publications, Inc.

Eriksson, I., Unden, A., & Elofsson, S. (2001). Self-rated health. Comparisons between three different measures. Results from a population study. *International Journal of Epidemiology, 30,* 326-333.

Fox, J. (1991). *Regression diagnostics: an introduction*. 1st ed. Sage Publications, Inc.

Glasgow, N. (2004). Healthy aging in rural America. In (L. W. M. N. Glasgow, N. E. Johnson, Eds.) *Critical issues in rural health* (pp. 271-281). Ames, Iowa: Blackwell Publishing Professional.

Glasgow, N., Min, H., & Brown, D. (2013). Volunteerism of older in-migrants and long-term residents in rural retirement destinations. In N. Glasgow & E. H. Berry (Ed.), *Rural aging in 21st century America* (pp. 231-250). New York: Springer Publishing Company.

Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. 1st ed. Cengage Learning.

Hamilton, L. C. (1995). *Data analysis for social scientists*. 1st ed. Boston, MA: Duxbury Press.

Hawkes, R. K. (1971). The multivariate analysis of ordinal measures. *The American Journal of Sociology, 76*(5), 908-926.

Idler, E. L. & Angel, R. J. (1990). Self-rated health and mortality in the NHANES-1 epidemiologic follow-up study. *American Journal of Public Health, 80*, 446-452.

Kawachi, I., Kennedy, B. P., & Glass, R. (1999). Social capital and self-rated health: A contextual analysis. *American Journal of Public Health, 89*(8), 1187-1193.

Kennedy, B. P., Kawachi, I., Glass, R., & Prothrow-Stith, D. (1998). Income distribution, socioeconomic status, and self rated health in the United States: Multilevel analysis. *BMJ: British Medical Journal, 317*, 917-921.

Long, S. J., & Freese, J. (2003). *Regression models for categorical dependent variables using STATA*. A Stata Press Publication. STATA Corporation. College Station: TX.

Manor, O., Matthew, S., & Power, C. (2000). Dichotomous or categorical response? Analysing self-rated health and lifetime social class. *International Journal of Epidemiology, 29*, 149-157.

Mckelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

McMurdo, M. E. T. (2000). A healthy old age: Realistic or futile goal? *BMJ: British Medical Journal, 321*, 1149-1151.

Menard, S. (2001). *Applied logistic regression analysis*. 2nd ed. Sage Publications**,** Inc.

Miilunpalo, S., Vuori, I., Oja, P., Pasanen, M., & Urponen, H. (1997). Self-rated health status as a health measure: The predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. *Journal of Clinical Epidemiology, 50*(5), 517-528.

Mossey, J. M., & Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. *American Joumal of Public Health, 72,* 800-808.

National Center for Health Statistics. (2007). *Health, United States, 2007 with chartbook on trends in the health of Americans*. Hyattsville, MD.

O'Brien, R.M. (1982). Using rank-order measures to represent continuous variables. *Social Forces, 61*, 144-155.

Pohlmann, J. T., & Leitner, W.W. (2003). A comparison of ordinary least squares and logisitc regression. *Ohio Journal of Science*, 103(5), 118-125.

Reynolds, H. T. (1973). On "the multivariate analysis of ordinal measures". *American Journal of Sociology, 78*(6), 1513-1516.

Row, J. W., & Kahn, R. L. (1987). Human aging: Usual and successful. *Science, 273*, 143-149.

SMS Research & Marketing Services, Inc. (2006). *HHS update: Hawaii department of health, office of health status monitoring. Hawaii Health Survey, 2004, Procedure Manual*. Honolulu, HI.

Wardle, J., & Steptoe, A. (2003). Socioeconomic differences in attitudes and beliefs about healthy lifestyles. *Journal of Epidemiology & Community Health, 57*, 440-443.

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review, 49*, 512-525.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 381-404.

ISSN 1538 − 9472

On Comparison of Exponential and Hyperbolic Exponential Growth Models in Height/Diameter Increment of PINES (*Pinus caribaea)*

Oyamakin S. O.
University of Ibadan
Ibadan, Nigeria

Chukwu A. U.
University of Ibadan
Ibdan, Nigeria

Bamiduro T. A.
Redeemer's University
Ogun State, Nigeria

A new tree growth model called the hyperbolic exponential nonlinear growth model is suggested. Its ability in model prediction was compared with the Malthus or exponential growth model an approach which mimicked the natural variability of heights/diameter increment with respect to age and therefore provides more realistic height/diameter predictions as demonstrated by the results of the Kolmogorov Smirnov test and Shapiro-Wilk test. The mean function of top height/Dbh over age using the two models under study predicted closely the observed values of top height/Dbh in the Hyperbolic exponential nonlinear growth models better than the ordinary exponential growth model without violating most of the assumptions about the error term.

Keywords: Model, height, Dbh, forest, Pinus caribaea, hyperbolic.

Introduction

The Caribbean Pine, *Pinus caribaea*, is a hard pine, native to Central America, Cuba, the Bahamas, and the Turks and Caicos Islands. It belongs to *Australes* Subsection in *Pinus* Subgenus. It inhabits tropical and subtropical coniferous forests, which include both lowland savannas and montane forests. Wildfire plays a major role limiting the range of this species, but it has been reported that this tree regenerates quickly and aggressively, replacing latifoliate trees. In zones not subject to periodic fires, the succession continues and a tropical forest thrives. It has been widely cultivated outside its natural range, and introduced populations can be found today in Jamaica, Colombia, South Africa or China. The species has three distinct varieties, one very distinct and treated as a

Oyamakin Oluwafemi Samuel is a Lecturer in the Deptartment of Statistics. Email at: fm_oyamakin@yahoo.com. Chukwu Angela Unna is in the Department of Statistics. Bamiduro Timothy Adebayo is in the Department of Mathematical Sciences.

separate species by some authors. These are *Pinus caribaea* var. *caribaea*, *Pinus caribaea* var. *bahamensis* (Bahamas Pine), and *Pinus caribaea* var. *hondurensis* (Honduras Pine).

Pines are a member of the gymnosperms, which literally means ‘naked seed’. This is because the ovule (which develops into the seed) is not enclosed during fertilization within a fruit-like structure like it is in flowering plants. Gymnosperms are an ancient lineage of plants that were abundant during the era of the dinosaurs. Pines are wind ‘pollinated’ and do not have flowers. They bear their seeds in distinctive pinecone. Other gymnosperms in Belize include the cycads that are common in the savanna and mountain cypress (*Podocarpus guatemalensis*) a tree found particularly in upland forests.

Figure 1. Growing Pines

Figure 2. A young Pine

A mathematical description of a real world system is often referred to as a mathematical model. A system can be formally defined as a set of elements also called components. A set of trees in a forest stand, producers and consumers in an economic system are examples of components. The elements (components) have certain characteristics or attributes and these attributes have numerical or logical values. Among the elements, relationships exist and consequently the elements are interacting. The state of a system is determined by the numerical or logical values of the attributes of the system elements. Experimenting on the state of a system with a model over time is termed simulation (Kleijnen, 1987). Scientific forest management relies to a large measure on the predictions of the future

conditions of individual stands. This is achieved by predicting the increment from the current stand structure and updating the current values at each cycle of iteration using a growth model. The structural changes over time can be monitored under different cutting cycles and cutting intensities and optimal management policies can be arrived at based on the results of such simulation runs

Jayaraman and Bailey (1988) proposed a growth model useful for simulating the changes occurring in an uneven aged mixed species stand. The mean annual increment in basal area and number of trees is predicted from the current values of basal area, number of trees, site quality and species composition of the stand and the simulation proceeds by progressive updating of the values of predictor variables in annual cycles. Changes in site quality are carried forward through a linear difference equation. Volume estimates at each time point can be obtained by an appropriate height-diameter relation and a volume table function.

Kumar (1988) reviews the different supply and demand models available in forestry and suggests a new model for a small wood producing country. The model essentially consists of a supply equation, an export function, a home demand equation and ar: identity on the inventories. Functional forms for the equations will have to be determined by empirical verification. Parameters can be estimated if data are available on a lengthy time series basis after converting the model to its reduced form. The reduced form expresses each current exogenous variable as a function of exogenous and lagged endogenous variables. Deterministic simulation can then be undertaken by tracing the time path of endogenous variables by specifying initial values for exogenous and lagged endogenous variables.

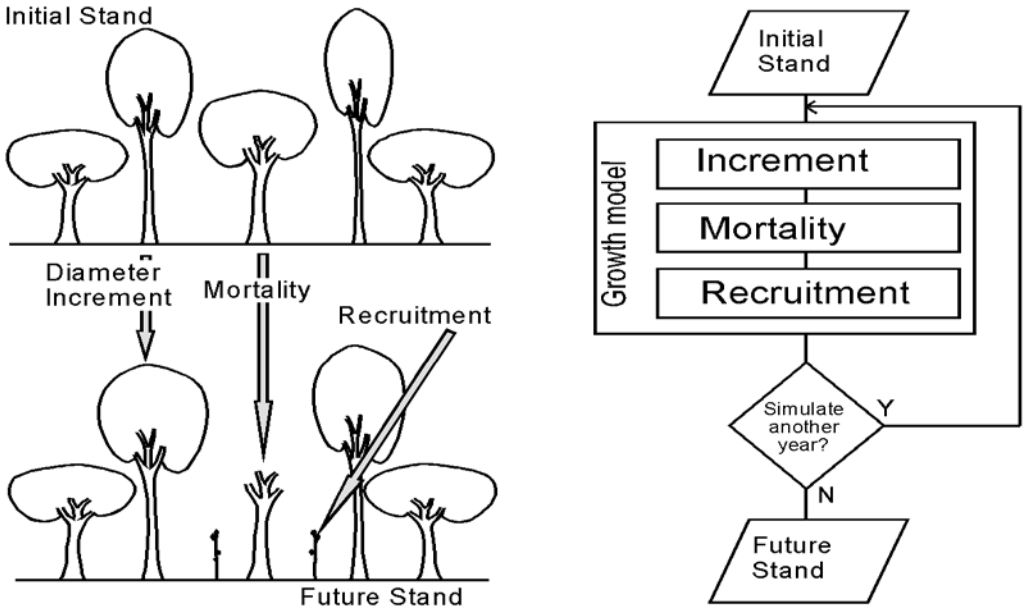


Figure 3. Components of forest growth and the analogous representation in a stand growth model.

Growth models assist forest researchers and managers in many ways. Some important uses include the ability to predict future yields and to explore silvicultural options. Models provide an efficient way to prepare resource forecasts, but a more important role may be their ability to explore management options and silvicultural alternatives. For example, foresters may wish to know the long-term effect on both the forest and on future harvests, of a particular silvicultural decision, such as changing the cutting limits for harvesting. With a growth model, they can examine the likely outcomes; both with the intended and alternative cutting limits and can make their decision objectively. The process of developing a growth model may also offer interesting new insights into stand dynamics.

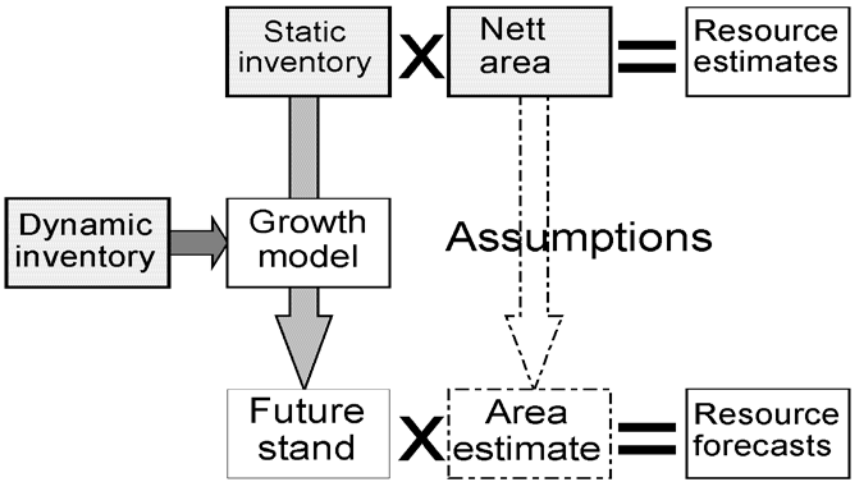


Figure 4. The role of growth models and complementary data in providing forest management information.

The total height (*Ht*) of a tree is important for assessing tree volume (Walters et al., 1985; Walters and Hann, 1986) and stand productivity through site index (Hann and Scrivani, 1987), but accurate measurement of this variable is time consuming. As a result, foresters often choose to measure only a few trees' heights and estimate the remaining heights with height-diameter equations. Foresters can also use height-diameter equations to indirectly estimate height growth by applying the equations to a sequence of diameters that were either measured directly in a continuous inventory or predicted indirectly by a diameter-growth equation. The diameter-growth prediction approach can be valuable for modeling growth and yield of trees and stands as it's done in ORGANON (Hann et al., 1997). A number of studies of height-diameter relationships in northwestern Oregon, western Washington, and southwest British Columbia have already been published. Curtis (1967) investigated several equations for Douglas-fir that included tree diameter outside bark at breast height (*DBH*) as an explanatory variable. Larsen and Hann (1987), and Wang and Hann (1988), using a variant of Curtis's (1967) recommended model, found that an equation which included tree diameter and site index was a better height predictor for 6 of 16 species in the mid-Willamette Valley. Krumland and Wensel (1988) included top height and quadratic mean diameter in their height-diameter equation.

Predicting total tree height based on observed diameter at breast height outside bark is routinely required in practical management and silvicultural

research work (Meyer, 1940). The estimation of tree volume, as well as the description of stands and their development over time, relies heavily on accurate height-diameter functions (Curtis, 1967). Many growth and yield models also require height and diameter as two basic input variables, with all or part of the tree height predicted from measured diameters (Burkhart et al., 1972; Curtis et al., 1981; Wykoff et al., 1982). In the cases where actual measurements of height growth are not available, height-diameter functions can also be used to indirectly predict height growth (Larsen and Hann 1987). Curtis (1967) summarized a large number of available height-diameter functions and used Furnival's index of fit to compare the performance of 13 linear functions fitted to second-growth Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) data. Since then, many new height-diameter functions have been developed. With the relative ease of fitting nonlinear functions and the nonlinear nature of the height-diameter relationships, nonlinear height-diameter functions have now been widely used in height predictions (Schreuder et al., 1979; Curtis et al., 1981; Wykoff et al., 1982; Wang and Hann 1988; Farr et al., 1989; Arabatzis and Burkhart, 1992).

Individual tree heights and diameters are essential measurements in forest inventories, and are used in estimating timber volume, site index and other important variables related to forest growth and yield, succession and carbon budget models (Peng, 2001). The time taken to measure tree heights takes longer than measuring the diameter at breast height. For this reason, often only the heights of a subset of trees of known diameter are measured, and accurate height-diameter equations must be used to predict the heights of the remaining trees to reduce the cost involved in data acquisition. If stand conditions vary greatly within a forest, a height regression may be derived separately for each stand, or a generalized function, which includes stand variables to account for the variability, may be developed (Curtis, 1967; Zhang et al., 1997; Sharma and Zhang, 2004). Two trees within the same stand and that have the same diameter are not necessarily of the same height; therefore a deterministic model does not seem appropriate for mimicking the real natural variability in height (Parresol and Lloyd, 2004).

The objective of the present study was to evaluate the performance of a stochastic height-diameter approach in mimicking the observed natural variability in *Gmelina Arborea* heights recorded in 2011.

Material and Methods

A fundamental nonlinear least squares assumption is that the error term in all the height-diameter functions considered are independent and identically distributed with zero mean and constant variance. However, in many forestry situations there is a common pattern of increasing variation as values of the dependent variable increase. This is clearly evident from the scatterplots of height versus *DBH* in Figure 2, where the values of the error are more likely to be small for small *DBH* and large for large *DBH*. When the problem of unequal error variances occurs, weighted nonlinear least squares (WNLS) is applied, with the weights selected to be inversely proportional to the variance of the error terms.

We used data from *Gmelina Arborea* even-aged stands located in Federal College of Forestry, Ibadan. The stand conditions within the plantation were similar and thus we consider the data obtained as belonging to the stands.

Method of Estimation

Consider a nonlinear model

$$H_i = f\left(D_i, \boldsymbol{B}\right) + \boldsymbol{\mathcal{E}_i} \tag{1}$$

$i = 1, 2, \ldots, n$, Where H is the response variable, D is the independent variable, $\boldsymbol{B}$ is the vector of the parameters β_j to be estimated ($\beta_1, \dagger \beta_2 \ldots\ldots., \beta_p$), $\mathcal{E}_i$ is a random error term , p is the number of unknown parameters, n is the number of observation. The estimator of β_j 's are found by minimising the sum of squares residual (SS_{Rss}) function

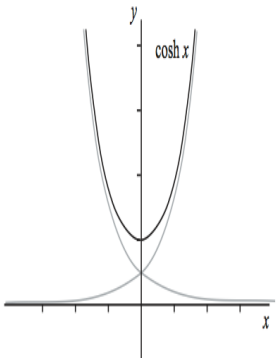
$$SS_{Rss} = \sum_{i=1}^{n} \left[H_i - f\left(D_i, B\right) \right]^2 \tag{2}$$

Under the assumption that the $\boldsymbol{\mathcal{E}_i}$ are normal and independent with mean zero and common variable σ^2 . Since H_i and D_i are fixed observations, the sum of squares residual is a function of $\boldsymbol{B}$, these normal equations take the form of

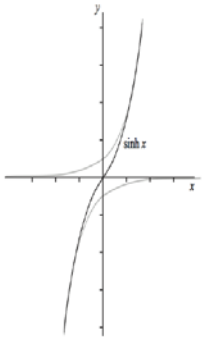
$$\sum_{i=1}^{n}\left\{H_i - f\left(D_i, \boldsymbol{B}\right)\right\}\left[\frac{\partial f\left(D_i, B\right)}{\partial \beta_j}\right] = 0 \qquad (3)$$

For $j = 1, 2, \ldots, p$. When the model is nonlinear in the parameters so are the normal equations consequently, for the nonlinear model, consider Table 2, it is impossible to obtain the closed solution of the least squares estimate of the parameter by solving the p normal equations describe in Eq (3). Hence an iterative method must be employed to minimize the ss_{Res} (Draper and Smith 1981, Ratkowsky 1983).

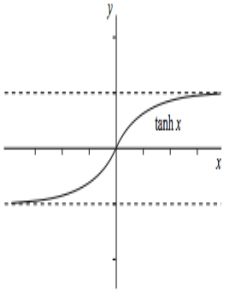
The hyperbolic functions have similar names to the trigonometric functions, but they are defined in terms of the exponential function. The three main types of hyperbolic functions and the sketch of their graphs are given below.



(a) Cosh Function



(b) Sinh function



(c) Tanh Function

The function (b) above is pronounced as ‘shine’, or sometimes as ‘sinch’. The function is defined by the formula

$$\sinh x = \frac{e^{x} - e^{-x}}{2}$$

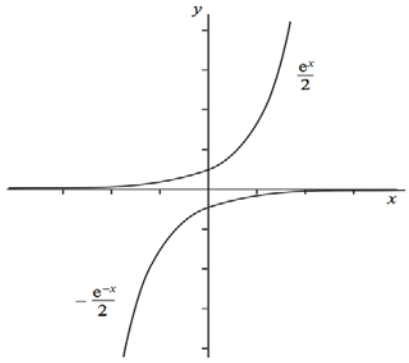
Again, we can use our knowledge of the graphs of *ex* and *e*−*x* to sketch the graph of sinh *x*. First, let us calculate the value of sinh 0. When $x = 0$, $e^{x} = 1$ and $e^{-x} = 1$. So

$$\sinh x = \frac{e^0 - e^{-0}}{2} = \frac{1-1}{2} = 0$$

Next, let us see what happens as *x* gets large. We shall rewrite sinh *x* as;

$$\sinh x = \frac{e^x}{2} - \frac{e^{-x}}{2}$$

To see how this behaves as x gets large, recall the graphs of the two exponential functions.



Graph of exponential functions

As *x* gets larger, e^x increases quickly, but e^{-x} decreases quickly. So the second part of the difference $\frac{e^x}{2} - \frac{e^{-x}}{2}$ gets very small as *x* gets large. Therefore, as *x* gets larger, sinh *x* gets closer and closer to $\frac{e^x}{2}$. This is written as;

$$\text{Sinh}x \approx \frac{e^x}{2} \text{ For large } x$$

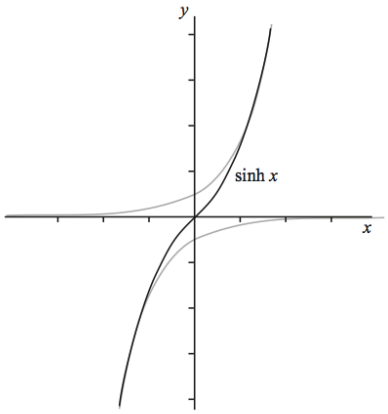
But the graph of Sinh *x* will always stay below the graph $\frac{e^x}{2}$. This is because, even though $-\frac{e^{-x}}{2}$ (the second part of the difference) gets very small, it is always less than zero. As *x* gets larger and larger the difference between the two graphs gets smaller and smaller.

Next, suppose that *x* is negative. As becomes more negative, $-e^{-x}$ becomes large and negative very quickly, but *ex* decreases very quickly. So as *x* becomes more negative, the first part of the difference $\frac{e^x}{2}-\frac{e^{-x}}{2}$ gets very small. So sinh *x* gets closer and closer to $-\frac{e^{-x}}{2}$. This is written as;

$\text{Sinh}x \approx \frac{-e^x}{2}$ For large negative *x*

Now the graph of sinh *x* will always stay above the graph of $\frac{e^{-x}}{2}$ when *x* is negative. This is because, even though $\frac{e^x}{2}$ (the first part of the difference) gets very small, it is always greater than zero. But as *x* gets more and more negative the difference between the two graphs gets smaller and smaller.

We can now sketch the graph of sinh *x*. Notice that sinh(−*x*) = − sinh *x*.



Graph of sinh (*x*)

Hence, the hyperbolic sine function and its inverse provide an alternative method for evaluating;

$$\int \frac{1}{\sqrt{1+x^2}}\,dx$$

If we make the substitution, then;

$$\sqrt{1+x^2} = \sqrt{1+sinh^2(u)} = \sqrt{cosh^2(u)} = \cosh(u)$$

Where the second equality follows from the identity $\cosh^2(u) - \sinh^2(u) = 1$ and the last equality from the fact that $\cosh(u) > 0$ for all u. Hence;

$$\int \frac{1}{\sqrt{1+x^2}}\,dx = \int \frac{\cosh(u)}{\cosh(u)}\,du = \int du = u + c = sinh^{-1}(x) + c$$

The following proposition is a consequence of the integral above i.e.

$$\frac{d}{dx} sinh^{-1}(x) = \frac{1}{\sqrt{1+x^2}}$$

Also, using the substitution $x = \tan(u)$, $-\frac{\pi}{2} < u < \frac{\pi}{2}$, that

$$\int \frac{1}{\sqrt{1+x^2}}\,dx = log\left|x + \sqrt{1+x^2}\right| + c$$

Since two anti-derivatives of a function can differ at most by a constant, there must exist a constant k such that

$$sinh^{-1}(x) = log\left|x + \sqrt{1+x^2}\right| + k$$

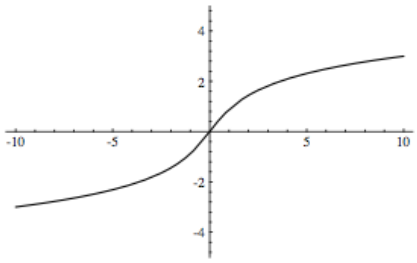
for all x. Evaluating both sides of this equality at $x = 0$, we have

$$0 = sinh^{-1}(0) = \log(1) + k = k$$

Thus $k = 0$ and

$$sinh^{-1}(x) = log\left|x + \sqrt{1 + x^2}\right|$$

for all x. Since the hyperbolic sine function is defined in terms of the exponential function, we should not find it surprising that the inverse hyperbolic sine function may be expressed in terms of the natural logarithm function.



Graph of arcsinh (*x*)

Hyperboloastic Exponential Growth Model (HEGM)

$$\frac{\partial H}{\partial t} = H\left[r + \frac{\theta}{\sqrt{1+t^2}}\right]$$

Separating the variables we have that;

$$\frac{\partial H}{H} = \left[r + \frac{\theta}{\sqrt{1+t^2}}\right]dt$$

Integrating both sides we have that;

$$\ln H = rt + \theta \operatorname{arcsinh}(t) + C_1$$

Hence,

$$H = Ae^{rt+\theta \text{arcsinh}(t)}$$

Therefore, we shall apply the two models below on Age-height and Age-Diameter of pines (*pinus carean*) growth;

(1) $H = Ae^{rt+\theta \text{arcsinh}(t)} + \varepsilon$, and $D = Ae^{rt+\theta \text{arcsinh}(t)} + \varepsilon$

(2) $H = Ae^{rt} + \varepsilon$, and $D = Ae^{rt} + \varepsilon$

Result and Discussion

Tables 1-4 below shows the estimated parameter for exponential and hyperbolic exponential growth model with their respective coefficient of determination (R^2) for age-height/age-diameter models

Table 1. Height Parameter Estimates using Exponential growth model

			95% Confidence Interval	
Parameter	**Estimate**	**Std. Error**	**Lower Bound**	**Upper Bound**
a	9.33	0.559	8.138	10.522
b	0.013	0.001	0.01	0.015

***R*-Square = 90.9%**

Table 2. Height Parameter Estimates using Hyperbolic Exponential growth model

			95% Confidence Interval	
Parameter	**Estimate**	**Std. Error**	**Lower Bound**	**Upper Bound**
a	2.178	.992	.051	4.306
b	.001	.003	-.006	.009
c	.448	.138	.153	.743

***R*-Square = 95.2%**

Table 3. Diameter Parameter Estimates using Exponential growth model

			95% Confidence Interval	
Parameter	**Estimate**	**Std. Error**	**Lower Bound**	**Upper Bound**
a	10.945	.515	9.847	12.043
b	.013	.001	.011	.015

***R*-Square = 94.5%**

Table 4. Diameter Parameter Estimates using Hyperbolic Exponential growth model

			95% Confidence Interval	
Parameter	**Estimate**	**Std. Error**	**Lower Bound**	**Upper Bound**
a	2.503	.680	1.044	3.963
b	.002	.002	-.003	.006
c	.452	.082	.276	.628

***R*-Square = 98.3%**

Also, the predicted and observed height and diameter were plotted to show the relationship and how best the models predicted the observed data on height and diameter of pines. This is also shown in the figure below:

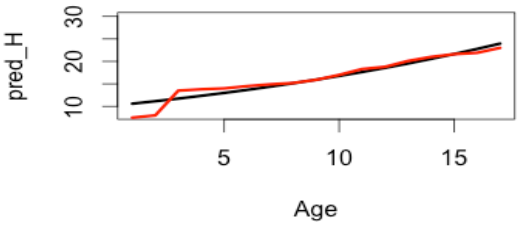


Figure 5. Observed Height against Predicted height
(Exponetial growth model)

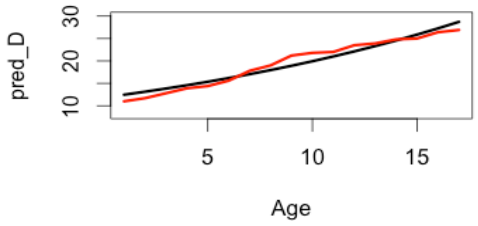


Figure 6. Observed Diameter against Predicted diameter
(Exponetial growth model)

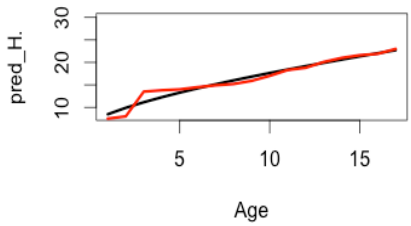


Figure 7. Observed Height against Predicted height
(Hyperbolic exponetial growth model)

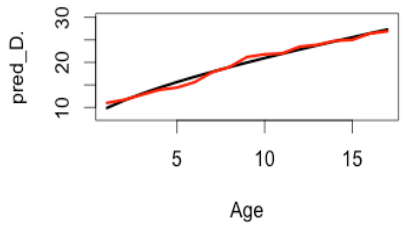


Figure 8. Observed Diameter against Predicted diameter
(Hyperbolic exponetial growth model)

Table 5. ANOVA summary for Height Parameter Estimates using Exponential growth

Source	Sum of Squares	*df*	Mean Squares
Regression	4873.136	2	2436.568
Residual	29.424	15	1.962
Uncorrected Total	4902.560	17	
Corrected Total	323.678	16	

Dependent variable: height
a. *R* squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .909.

Table 6. ANOVA summary for Height Parameter Estimates using Hyperbolic Exponential growth model

Source	Sum of Squares	*df*	Mean Squares
Regression	4886.955	3	1628.985
Residual	15.605	14	1.115
Uncorrected Total	4902.560	17	
Corrected Total	323.678	16	

Dependent variable: height
a. *R* squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .952.

Table 7. ANOVA summary for Diameter Parameter Estimates using Exponential growth model

Source	Sum of Squares	*df*	Mean Squares
Regression	6910.833	2	3455.417
Residual	25.417	15	1.694
Uncorrected Total	6936.250	17	
Corrected Total	464.198	16	

Dependent variable: height
a. *R* squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .945.

Table 8. ANOVA: Diameter Parameter Estimates using Hyperbolic Exponential growth model

Source	Sum of Squares	*df*	Mean Squares
Regression	6928.553	3	2309.518
Residual	7.697	14	.550
Uncorrected Total	6936.250	17	
Corrected Total	464.198	16	

Dependent variable: height
a. *R* squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .983.

Testing for Independence of Errors (Run test) and Normality of Error (Shapiro-Wilk test)

Two assumptions made in the models are:

Errors are independent
Errors are normally distributed.

These assumptions were verified by examining the residuals. If the fitted models are correct, residuals should exhibit tendencies that tend to confirm or at least should not exhibit a denial of the assumptions.

Hence, we tested the following hypotheses stated below;

H_0: Errors are independent (Using Runs Test)
H_1: Errors are not independent

And

H_0: Errors are normally distributed (Using Shapiro-Wilk test)
H_1: Errors are not normally distributed

Let m be the number of pluses and n be the number of minuses in the series of residuals. The test is based on the number of runs(r), where a run is defined as a sequence of symbols of one kind separated by symbols of another kind. A good large sample approximation to the sampling distribution of the number of runs is the normal distribution with mean;

$$Mean = \frac{2mn}{m+n} + 1$$

and,

$$Variance\left(\sigma^2\right) = \frac{2mn\left(2mn - m - n\right)}{(m+n)^2\left(m+n-1\right)}$$

Therefore, for large samples like ours the required test statistic is;

$$Z = \frac{\left(r + h - \mu\right)}{\sigma} \sim N\left(0,1\right)$$

where,

$$h = \begin{cases} 0.5, \ if \ \ r < \mu \\ -0.5, \ r > \mu \end{cases}$$

Also, the required test statistic for the test of normality (Shapiro-Wilk test) is given by;

$$W = \frac{S^2}{b}$$

Where;

$$S^2 = \sum a(k)\left\{x_{n+1-k} - x_{(k)}\right\}$$

and,

$$b = \sum \left(x_i - \bar{x}\right)^2$$

In the above, the parameter k takes the values; $x_{(k)}$ is the k^{th} order statistic of the set of residuals and the values of coefficient $a(k)$ for different values of n and k are given in the Shapiro-Wilk table (1965). H_0 is rejected at level α i.e. W is less than the tabulated value.

Table 9. Result of the test of independence of Residuals using Run Test

Residual	Test Value	No. of Runs	*Z*	Asymp. Sig.(2 tailed)
Exp. Height	-0.2000	5	-1.802	0.072*
Exp. Diameter	-0.0318	3	-3.002	0.003***
*H*Exp. Height	-0.0047	6	-1.494	0.135^{ns}
*H*Exp. Diameter	0.0035	4	-2.499	0.012**

* Significant at 10%, ** significant at 5%, *** significant at 99% and ^{ns} not significant

Table 10. Result of the test of normality of Residuals using K-S & S-W Tests

Residual	Kolmogorov-Sminov		Shapiro-Wilk	
	Statistic	Asmp. Sig.	Statistic	Asmp. Sig.
Exp. Height	0.262	***0.003******	0.842	***0.008******
Exp. Diameter	0.198	***0.077****	0.933	$***0.244^{ns}***$
H Exp. Height	0.172	$***0.193^{ns}***$	0.954	$***0.519^{ns}***$
H Exp. Diameter	0.192	$***0.095^{ns}***$	0.953	$***0.500^{ns}***$

* Significant at 10%, ** significant at 5%, *** significant at 99% and ^{ns} not significant

Conclusion

The mean function of top height and *Dbh* over age using the Hyperbolic Exponential growth model predicted closely the observed values of top height and

Diameter of Pines. However, large correlations of the estimated parameters do not necessary mean that the original model is inappropriate for the physical situation under study. For example, in a linear model, when a particular β (a coefficient) does not appear to be different from zero, it does not always imply that the corresponding x (independent variable) is ineffective; it may be that, in a particular set of data under study, x does not change enough for its effect to be discernible. In general, efficient parameter estimation can best be achieved through a good understanding of the meaning of the parameters, the mathematics of the model, including the partial derivatives, and the system being modeled.

References

Arabatzis, A. A., & Burkhart, H. E. (1992). An evaluation of sampling methods and model forms for estimating height-diameter relationships in a loblolly pine plantation. *For. Sci, 38*, 192-198.

Brisbin, I. L. (1989). Growth curve analyses and their applications to the conservation and captive management of crocodilians. *Proceedings of the Ninth Working Meeting of the Crocodile Specialist Group.* Gland, Switzerland: SSCHUSN.

Burkhart, H. E., Parker, R. C., Strub, M. R., & Oderwald R. G. (1972). *Yield of old-field loblolly pine plantations.* Blacksburg, VA: School of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University. Publ-FWS-3-72.

Bursac, Z, Tabatabai, M., & Williams, D. K. (2007). Nonlinear Hyperbolastic Growth Models and Application in Craniofacial and Stem Cell Growth. *ASA Biometrics Section (including ENAR and WNAR)* 190-197.

Castro, M. A., Klamt, F., Grieneisen, V., *et al*. (2003). Gompertzian growth pattern correlated with phenotypic organization of colon carcinoma, malignant glioma and non-small cell lung carcinoma cell lines. *Cell Prolif, 36*(2), 65-73.

Chignola, R., Schenetti, A., Chiesa, E., et al. (1999). Oscillating growth patterns of multicellular tumor spheroids. *Cell Prolif, 32*, 39-48.

Curtis, R. O. (1967). Height-diameter and height-diameter-age equations for second-growth Douglas-fir. *Forest science 13*, 365 – 375.

Curtis, R. O., Clendenen, G. W., & DeMars, D. J. (1981). *A new stand simulator for coast Douglas-fir: DFSIM Users guide.* USDA For. Serv. Gen. Tech. Rep. PNW – 128.

Deisboeck, T. S., Berens, M. E., Kansal, A. R., *et al.* (2001). Pattern of self-organization in tumour systems: complex growth dynamics in a novel brain tumour spheroid model. *Cell Prolif, 34*, 115-134.

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis.* New York: John Wiley and Sons.

Farr, W. A., DeMas, D. J., & Dealy, J. E. (1989). Height and crown width related to diameter for open-grown western hemlock and sitka spruce. *Can. J. For. Res. 19*, 1203 - 1207

Fekedulegn, D. (1996). *Theoretical nonlinear mathematical models in forest growth and yield modeling (Thesis).* Dept. of Crop Science, Horticulture and Forestry, University College Dublin, Ireland.

Foong, F. S. (1991). Potential evapotranspiration, potential yield and leaching losses of oil palm. *Pro. 1991 PORIM Interl. Palm Oil Cong. (Agric.),* 105 – 119.

Foong, F. S. (1999). Impact of moisture on potential evapotranspiration, growth and yield of palm oil. *Pro. 1999 PORIM Inter1.1 Palm oil Cong. (Agric),* 265 – 287

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Phil Trans of the Royal Soc, 182*, 513-585.

Hann, D. W., Hester, A. S., & Olsen, C. L. (1997). *Organon Users manual edition 6.0.* Department of Forest Resources, Oregon State University, Corvallis.

Hann, D. W., & Scrivani, J. (1987). *Dominant height-growth and site-index equations for Douglas-fir and ponderosa pine in southwest Oregon.* Research Bulletin 59, Department of Forest Resources, Oregon State University, Corvallis

Ismail, Z., Khamis, A., & Jaafar, M. Y. (2003). Fitting nonlinear Gompertz curve to tobacco growth data. *Pak. J. Agro., 2*, 223 – 236.

Jaafar, M. Y. (1999). Pensuaian model taklinear Gompertz terhadap pertumbesaran pokok koko. *Matematika, 15*, 1 – 20.

Jayaraman, K., & Bailey, R. L. (1988) Modelling mixed species stands with Scheffe's canonical polynomials. In Ek, A.R., et al. (Ed.) *Forest Growth Modelling and Prediction. Proceedings of the IUFRO Conference, Minnesota, 23-27 August 1987*, pp. 201-208. (KFRI Scientific Paper No. 515)

Kansal, A. R., Torquato, S., Harsh, G.R., *et al*. (2000). Simulated brain tumor growth dynamics using a three dimensional cellular automaton. *J Theor Biol, 203*, 367-82.

Khamis, A., & Ismail, Z. (2004). Comparative study on nonlinear growth curve to tobacco leaf growth data. *J. Agro., 3*, 147 – 153.

Kingland, S. (1982). The refractory model: The logistic curve and history of population ecology. *Quart Rev Biol, 57*, 29-51.

Kleijnen, J. P. C. (1987). Simulation with too many factors: review of random and group-screening designs. *European Journal of Operational Research 31*(1), 31-36.

Kramer, P. J., & Kozlowski, T. T. (1979). *Physiology of woody plants.* New York: Academic Press.

Kumar, M. (1988). Reed Bamboo (Ochlandra) in Kerala: Distribution and management. In Rao, I.V.R., et al. (Ed.). *Bamboos: Current Research. Proceedings of the International Bamboo Workshop, Cochin, 14-18 November 1988*, pp. 39-43. (KFRI Scientific Paper No. 34)

Kumar, M., & Manilal, K.S. (1988). Floral anatomy of Apostasia odorata and the taxonomic status of Apostasioids (Orchidaceae). *Phytomorphology, 38*(2/3), 159-162. (KFRI Scientific Paper No. 35)

Krumland, B. E., & Wensel, L. C. (1988). A generalized height-diameter equation for coastal California species. *Western Journal of Applied Forestry, 3*, 113 – 115

Larsen, D. R. & Hann, D. W. (1987). *Height-Diameter equations for seventeen tree species in southwest Oregon*. Research paper 49, Forest Research laboratory, Oregon State University, Corvallis.

Maraquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics, 11*, 431 – 441.

Martin, G. L., & Martin, A. R. (1984). A comparison of competition measures and growth models for predicting plantation red wine diameter and height growth. *Forest Science, 30*, 731-743.

Marusic, M., Bajzer, Z., Freyer, J. P., *et al.* (1994). Analysis of growth of multicellular tumor spheroids by mathematical models. *Cell Prolif, 27*, 73-94.

Marusic, M., Bajzer, Z., Vuk-Pavlovic, S., *et al.* (1994). Tumor growth in vivo and as multicellular spheroids compared by mathematical models. *Bull Math Biol, 56*, 617-31.

Methley, J. (1996). *Bowmont sample plot data-end user licensed for data release*. Forestry Commission Research Division, 12 June 1996, England.

Meyer, H. A. (1940). A mathematical expression for height curves. *Journal of Forestry 38,* 415 – 420.

Morgan, P. H., Mercer, L. P., & Flodin, N. W. (1975). General model for nutritional response of higher organisms. *Proc. Nat. Acad. Sci. USA, 72,* 4327 - 4331

Myers, R. H. (1996). *Classical and modern regression with applications*. Boston, MA: Duxbury Press.

Nelder, J. A. (1961). The fitting of a generalization of the logistic curve. *Biometrics, 17*, 89-110.

Olea, N., Villalobos, M., Nunez, M. I., *et al.* (1994). Evaluation of the growth rate of MCF-7 breast cancer multicellular spheroids using three mathematical models. *Cell Prolif, 27,* 213-23.

Oliver, F. R. (1964). Methods of estimating the logistic function. A*pplied statistics, 13*, 57-66.

Parresol, B. R., & Lloyd, F. T. (2004). *The stochastic tree modelling approach used to derive tree lists for the GIS/CISC identified stands at the Savannah River Site,* Internal Report, USDA For. Serv. Southern Research Station, Asheville, NC.

Paul, J. R. (1971). *History of Poliomyelitis.* London: Yale University Press.

Peng, M. (2001). The resource-based view and international business. *Journal of Management, 26*(3), 513 -563.

Philip, M. S. (1994). *Measuring trees and forests*. 2nd Edition. Wallingford, UK: CAB International.

Phillips, B. F., & Campbell, N. A. (1968). A new method of fitting the von Bertalanffy growth curve using data on the whelk. *Dicathais Growth 32,* 317-329.

Ratkowskay, D. A. (1983). *Nonlinear Regression modeling*. New York: Marcel Dekker.

Richards, F. J. (1959). A flexible growth functions for empirical use. *Journal of Experimental Botany, 10,* 290-300.

Ricklef, R. E., & Scheuerlein, A. (2002). Biological implications of the Weibull and Gompertz models of aging. *J Gerotol A Biol Sci Med Sci*, *57*(2), B69-B76.

SAS Institute Inc. (1985). *SAS/STAT User's Guide, version 6*, 4th edition. Vol. 1. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1992). *SAS/STAT: User's Guide, Release 6.03 edition.* Cary, NC: SAS Institute Inc.

Schnute, J. (1981). A versatile growth model with statistically stable parameters. *Can. J. Fish. Aquat. Sci. 38*, 1128.

Schreuder, H. T., Hafley, W. L., & Bennett, F. A. (1979). Yield prediction for unthinned natural slash pine stands. *For. Sci 25*, 25 – 30.

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear Regression*. New York: John Wiley and Sons.

Sharma, M., & Zhang, S. Y. (2004). Height-Diameter models using stand characteristics for pinus banksiana and picea mariana. *Scandinavian Journal of Forest Research, 19*, 442-451.

Spratt, J. A., von Fournier, D., Spratt, J. S., *et al.* (1993). Decelerating growth and human breast cancer. *Cancer, 71*, 2013-9.

Tabatabai, M., Williams, D. K., & Bursac, Z. (2005). Hyperbolastic growth models: Theory and application. *Theor Biol Med Model, 2*(14), 1-13.

Tsoularis, A., & Wallace, J. (2002). Analysis of logistic growth models. Math. *Biosci., 179*, 21 – 55.

Vanclay, J. K. (1994). *Modeling forest growth and yield.* Wallingford, UK: CAB International.

Verhulst, P. F. (1838). A note on population growth. *Correspondence Mathematiques et Physiques, 10,* 113-121.

Von Bertalanffy, L. (1957). Quantitative laws in metabolism and growth. *Quantitative Rev. Biology 32*, 218-231.

Walters, D. K., Hann, D. W., & Clyde, M. A. (1985). *Equations and tables predicting gross total stem volume in cubic feet for six major conifers of southwest Oregon*. Research Bulletin 50, Forest Resources, Oregon State University, Corvallis.

Wang, C. H., & Hann, D. W. (1988). *Height-Diameter equations for sixteen tree species in the central western Willamette valley of Oregon,* Research paper 51, Forest Research Laboratory, Oregon State University, Corvallis.

Walters, D. K., & Hann, D. W. (1986). *Predicting merchantable volume in cubic feet to a variable top and in scribner board feet to a 6-inch top for six major conifers of southwest Oregon*, Research Bulletin 52, Forest Resources, Oregon State University, Corvallis.

Weibull, W. (1951). A statistical distribution function of wide applicability. *J of Appl Mech, 18*, 293-297.

West, G. B., Brown, J. H., & Enquist, B. J. (2001). A general model for antogenetic growth. *Nature, 413*, 628-631.

West, G. B., Brown, J. H.,, & Enquist, B. J. (2004). Growth models based on first principles of Phenomenology? *Funct Ecol, 18*, 188-196.

Wykoff, W. R., Crookston, N. L. & Stage, A. R. (1982). *User's guide to the stand prognosis model*. General Technical Report INT-133, USDA Forest Service, Inter-Mountain Forest and Range Experiment Station, Ogden UT.

Yin, X., Goudriaan, J., Latinga, E.A., et al. (2003). A flexible sigmoid function of determinate growth. *Ann Bot (Lond), 91*, 361-371.

Yuancai, L., Marques, C. P., & Macedo, F. W. (1997). Comparison of Schnute's and Bertalanffy-Richards' growth function. *Forest Eco. Mgmt., 96*, 283 – 288.

Zeide, B. (1993). Analysis of growth equations. *Forest Sci, 39*, 594-616.

Zhang, L. (1997). Cross-validation of non-linear growth functions for modeling tree height-diameter relationships. *Annals of Botany, 79*, 251-257.

Zhu, Q., Cao, X., & Luo, Y. (1988). Growth analysis on the process of grain filling in rice. *Acta Agronomica Sinica, 18*, 182-193.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 405-426.

ISSN 1538 − 9472

Approximation Multivariate Distribution of Main Indices of Tehran Stock Exchange with Pair-Copula

G. Parham
Shahid Chamran University
Ahvaz, Iran

A. Daneshkhah
Cranfield University
Cranfield, UK

O. Chatrabgoun
Shahid Chamran University
Ahvaz, Iran

The multivariate distribution of five main indices of Tehran stock exchange is approximated using a pair-copula model. A vine graphical model is used to produce an n-dimensional copula. This is accomplished using a flexible copula called a minimum information (MI) copula as a part of pair-copula construction. Obtained results show that the achieved model has a good level of approximation.

Keywords: Minimum information copula, pair-copula, vine.

Introduction

Sometimes in applied probability and statistics it is necessary to model multiple uncertainties or dependencies using multivariate distributions. To do it, it is common to use discrete model such as Bayesian networks but when modeling financial data, it is necessary to have model of continuous random variables. Copulas are quickly gaining popularity as modeling dependencies e.g. surveys by Nelsen (1999), Joe (1997). Copulas have found application in a number of areas of operations research including combining expert opinion and stochastic simulation, (e.g. Abbas et al. (2010) and references cited therein). A copula is a joint distribution on the unit square (or more generally on the unit n-cube) with uniform marginal distributions. Under reasonable conditions, a joint distribution for n-random variables can be found by specifying the univariate distribution for each variable, and in addition, specifying the copula. Following Sklar (1959) the joint distribution function of random vector $\left(X_1,\ldots,X_n\right)$ is

Dr. Parham is Faculty of Mathematical Sciences and Computer. Email at: Parham_g@scu.ac.ir. Dr. Daneshkhah is a Lecturer in Utility Asset Management. Email at: ardaneshkhah@gmail.com. Dr. Chatrabgoun is Faculty of Mathematical Sciences and Computer. Email at: o-chatrabgoun@phdstu.scu.ac.ir.

$$F\left(x_1,\ldots,x_n\right)=C\left(F_1\left(x_1\right),\ldots,F_n\left(x_n\right)\right) \quad (1)$$

Where C is a copula distribution function, and $F_1,\ldots,F_n$ are the univariate, or marginal, distribution functions. A special case is that of the 'Gaussian copula', obtained from Gaussian joint distribution and parameterized by the correlation matrix. Use of the Gaussian copula to construct joint distributions is equivalent to the NORTA method (normal to anything). Clearly the use of a copula to model dependency is simply a translation of one difficult problem into another: instead of the difficulty of specifying the full joint distribution is the difficulty of specifying the copula. The main advantage is the technical one that copulas are normalized to have support on the unit square and uniform marginals. As many authors restrict the copulas to a particular parametric class (Gaussian, multivariate t, etc.) the potential flexibility of the copula approach is not realized in practice.

As mentioned because of difficulty in specifying the copulas and restricted to the exact class, copula approximation is to some extend new topic in this case. The approach used herein allows a lot of flexibility in copula specification that was analyzed and some properties of it was said in Bedford et al. (2013) and developed by Daneshkhah et al. (2013), and for approximation multivariate distribution, a graphical model, called a vine, is used to systematically specify how two-dimensional copulas are stacked together to produce an n-dimensional copula.

The main objectives is to show that a vine structure can be used to approximate Tehran stock exchange multivariate copula to any required degree of approximation. The standing technical assumptions are that the multivariate copula density f under study is continuous and is non-zero. No other assumptions are needed. A constructive approach involves the use of minimum information (*MI*) copula that can be specified to any required degree of precision based on the data available. According to Bedford et al. (2013) good approximation locally guarantees good approximation globally.

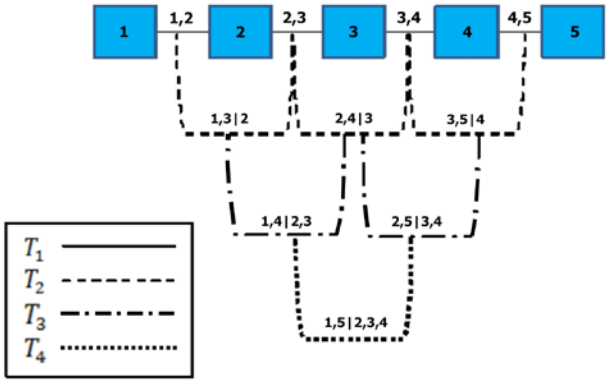


Figure 1. A Regular Vine with 5 Elements

A vine structure imposes no restrictions on the underlying joint probability distribution it represents (as opposed to the situation for Bayesian networks, for example). However this does not mean to ignore the question about which vine structure is most appropriate, for some structures allow the use of less complex conditional copulas than others. Conversely, if only certain families of copulas are allowed then one vine structure might fit better than another.

Vine constructions for multivariate dependency

A copula is a multivariate distribution function with standard uniform marginal distributions. Using (1) it may be observed that a copula can be used, in conjunction with the marginal distributions, to model any multivariate distribution. However, apart from the multivariate Gaussian, Student, and the exchangeable multivariate Archimedean copulas, the set of higher-dimensional copulas proposed in the literature is limited and is not rich enough to model all possible mutual dependencies amongst the n variants (see Kurowicka & Cooke, 2006 for details of these copulas). Hence it is necessary to consider more flexible constructions.

A flexible structure, here denoted the pair-copula construction or vine, allows for the free specification of (at least) $n(n-1)/2$ copulas between n variables. (Note that $n(n-1)/2$ is the number of entries above the diagonal of an $n \times n$ correlation matrix - though these are algebraically related so not completely free variables). This structure was originally proposed by Joe (1997), and later

reformulated and discussed in detail by Bedford and Cooke (2001, 2002), who considered simulation, information properties and the relationship to the multivariate normal distribution but who also considered a more general method called a Cantor tree construction. Kurowicka and Cooke (2006) considered simulation issues, and Aas et al. (2009) examined inference. The modeling scheme is based on a decomposition of a multivariate density into a set of bivariate copulas. The way these copulas are built up to give the overall joint distribution is determined through a structure called a vine, and can be easily visualized. A vine on n variables is a nested set of trees, where the edges of the tree j are the nodes of the tree $j+1$ (for $j=1,...,n-2$), and each tree has the maximum number of edges. For example, Figure 1 shows a vine with 5 variables which consists of four trees (T_1,T_2,T_3,T_4) with 4, 3, 2 and 1 edges, respectively. A regular vine on n variables is a vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node, for $j=1,...,n-2$. There are $n(n-1)/2$ edges in a regular vine on n variables. The formal definition is as follows.

Definition: (Vine, regular vine) V is a vine on n elements if

1. $V=(T_1,\ldots,T_{n-1})$.

2. T_1 is a connected tree with nodes $N_1=\{1,...,n\}$ and edges E_1; for $i=2,...,n-1, T_i$ is a connected tree with nodes $N_1 = E_{i-1}$.
V is a regular vine on n elements if additionally the proximity condition holds:

3. For $i=2,...,n-1$, if a and b are nodes of T_i connected by an edge in T_i, where $a=\{a_1,a_2\}, b=\{b_1,b_2\}$, then exactly one of the a_i equals one of the b_i.

One of the simplest regular vines is shown in Figure 1 - this structure is called D-vine, see Kurowicka and Cooke, 2006, pp. 93. Here, T_1 is the tree consisting of the straight edges between the numbered nodes. T_2 is the tree consisting of the curved edges that join the straight edges in T_1, and so on.

For a regular vine each edge of T_1 is labelled by two numbers from $\{1,...,n\}$. If two edges of T_1, for example 12 and 23, which are nodes joined by an edge in T_2 are taken, then of the numbers labeling these edges one is common to both (2), and they both have one unique number (1,3 respectively). The common number(s) will be called the conditioning set D_e for that edge e (in this example the conditioning set is simply {2}) and the other numbers will be called the conditioned set (in this example {1, 3}). For a regular vine the conditioned set always contains two elements.

A vine distribution is associated to a vine by specifying a copula to each edge of T_1 and a family of conditional copulas for the conditional variables given the conditioning variables, as shown by the following result of Bedford and Cooke (2001).

Theorem 1: Let $V=(T_1,...,T_{n-1})$ be a regular vine on n elements. For each edge $e(j,k)\in T_i$, $i=1,...,n-1$ with conditioned set $\{j,k\}$ and conditioning set D_e, let the conditional copula and copula density be $C_{jk|D_e}$ and $c_{jk|D_e}$ respectively. Let the marginal distributions F_i with densities f_i, $i=1,...,n$ be given. Then the vine-dependent distribution is uniquely determined and has a density given by

$$f(x_1,\ldots,x_n)=\prod_{i=1}^{n} f(x_i)\prod_{e(j,k)\in E_i} c_{jk|D_e}\left(F_{j|D_e(x_j)},F_{jk|D_e(x_k)}\right) \tag{2}$$

The existence of regular vine distributions is discussed in detail by Bedford and Cooke (2002).

The density decomposition associated with 5 random variables $X=(X_1,...,X_5)$ with a joint density function $f(x_1,\dagger\ldots,\dagger x_5)$ satisfying a copula-vine structure shown in Figure 1 with the marginal densities $f_1,\ldots,f_5$ is

$$\begin{aligned} f_{12345}=\prod_{i=1}^{n} f(x_i)\times\ & c_{12}\left(F(x_1),F(x_2)\right)c_{23}\left(F(x_2),F(x_3)\right)c_{34}\left(F(x_3),F(x_4)\right)c_{45}\left(F(x_4),F(x_5)\right)\\ & \times c_{13}\left(F(x_1\mid x_2),F(x_3\mid x_2)\right)c_{24}\left(F(x_2\mid x_3),F(x_4\mid x_3)\right)c_{35}\left(F(x_3\mid x_4),F(x_5\mid x_4)\right)\\ & \times c_{14}\left(F(x_1\mid x_2,x_3),F(x_4\mid x_2,x_3)\right)c_{25}\left(F(x_2\mid x_3,x_4),F(x_5\mid x_3,x_4)\right)\\ & \times c_{15}\left(F(x_1\mid x_2,x_3,x_4),F(x_5\mid x_2,x_3,x_4)\right) \end{aligned} \tag{3}$$

This formula can be derived for this case using the general expression

$$f_{12}(x,y) = f_1(x)\, f_2(y) c_{12}\left(F_1(x), F_2(y)\right)$$

or equivalently

$$f_{1|2}(x \mid y) = f_1(x) c_{12}\left(F_1(x), F_2(y)\right)$$

where c_{12} is the copula density and F_1, F_2 are the univariate distributions. Starting with

$$\begin{aligned} f_{12345}(x_1,\ldots,x_5) = {} & f_1(x_1)\, f_{2|1}(x_2 \mid x_1)\, f_{3|12}(x_3 \mid x_1, x_2) \\ & f_{4|123}(x_4 \mid x_1, x_2, x_3)\, f_{5|1234}(x_5 \mid x_1, x_2, x_3, x_4) \end{aligned}$$

inductively convert the latter expression in to that shown in (3). This results in

$$f_{2|1}(x_2 \mid x_1) = f_2(x_2) c_{12}\left(F_1(x_1), F_2(x_2)\right)$$

Next,

$$\begin{aligned} f_{3|12}(x_3 | x_1, x_2) &= f_{3|2}(x_3 | x_2) c_{13|2}\left(F_{1|2}(x_1 \mid x_2), F_{3|2}(x_3 \mid x_2)\right) \\ &= f_3(x_3) c_{23}\left(F_2(x_2), F_3(x_3)\right) c_{13|2}\left(F_1(x_1 \mid x_2), F_3(x_3 \mid x_2)\right) \end{aligned} \tag{4}$$

The calculation for the remaining term $f_{5|1234}(x_5 \mid x_1, x_2, x_3, x_4)$ is left to the reader.

Note that in the special case of a joint normal distribution, the normal copula would be used everywhere in the above expression and the conditional copulas would be constant (i.e. not depend on the conditioning variable). This means that the joint normal structure is specified by $n(n-1)/2$ (conditional) correlation values, which are algebraically free between -1 and +1 (unlike the values in a correlation matrix). See Bedford and Cooke (2002) for more details. The above theorem gives a constructive approach to build a multivariate distribution given a vine structure: If choices of marginal densities and copulas are made then the above formula will give a multivariate density. Hence, vines can be used to model general multivariate densities. However, in practice it is necessary to use copulas from a convenient class, and this class should ideally be one that allows any given

copula to be approximated to an arbitrary degree. Having this class of copulas allows any multivariate distribution to be approximated using any vine structure.

Unlike the situation with Bayesian networks, where not all structures can be used to model a given distribution, the theorem shows that - in principle - any vine structure may be used to model a given distribution. However, when specific families of copulas are used it seems that some vine structures do work better than others. That is, given a family of copulas, some vine structures may give a better degree of approximation than others. It is worth stressing the point that the flexibility of vines gives the potential to capture any fine grain structure within a multivariate distribution. A key aspect that cannot be modeled by Bayesian networks is that of conditional dependence. Bayesian networks are built around the concept of conditional independence -arrows from a parent node to two child nodes means that the child variables are conditionally independent given the parent variable. However, different models of conditional dependence are not available as building blocks in Bayesian networks. Multivariate Gaussian copulas do allow for a specification of conditional dependence, but do not allow that dependence to change - in a multivariate normal distribution, the conditional correlation of two variables given a third may be non-zero but is always constant. This approach, by contrast, allows the explicit modeling of non-constant conditional dependence.

The minimum information (MI) copula using the D_1AD_2 algorithm

Bedford et. al (2013) presented a way to approximate a copula using minimum information methods which demonstrate uniform approximation in the class of copula used. Bedford and Meeuwissen (1997) applied a so-called *DAD* algorithm to produce discretized MI copula with given rank correlation. This approach can be used whenever it is desirable to specify the expectation of any symmetric function of $U = F(x)$ and $V = F(y)$.

In order to have asymmetric specifications the D_1AD_2 algorithm must be used where A is a positive square matrix, thus, diagonal matrices D_1 and D_2 can be found such that the product of D_1AD_2 is doubly stochastic. It is possible to correlate the variables of interest X and Y by introducing constraints based on knowledge about functions of these variables. Suppose there are k of these functions, namely $h_1'(X,Y), h_2'(X,Y), \ldots, h_k'(X,Y)$ and mean values $\alpha_1, \ldots, \alpha_k$ are specified for all functions respectively from the data or the expert judgment.

Corresponding functions of the copula variables U and V, defined by $h_1(U,V) = h_1'\left(F_1^{-1}(U), F_2^{-1}(V)\right)$, etc. can be defined and clearly these should also have the specified expectations $\alpha_1,...,\alpha_k$. The kernel

$$A(u,v) = exp(\lambda_1 h_1(u,v) + \ldots + \lambda_k h_k(u,v)) \tag{5}$$

is formed, where u denotes the realization of U and v the realization of V.

For practical implementations it is necessary to discretize the set of (u,v) values such that the whole domain of the copula is covered. This means that the kernel A described above becomes a 2-dimensional matrix A and that the matrices D_1 and D_2 are required to create a discretized copula density

$$P = D_1 A D_2 \tag{6}$$

Suppose that both U and V are discretized into n points, respectively u_i, and v_j, $i, j = 1,...,n$. Then $A = \left(a_{ij}\right), D_1 = diag\left(d_1^{(1)},\ldots,d_n^{(1)}\right)$ and $D_2 = diag\left(d_1^{(2)},\ldots,d_n^{(2)}\right)$ where $a_{ij} = A\left(u_i, v_j\right), d_i^{(1)} = D_1\left(u_i\right)$ and $d_i^{(2)} = D_2\left(u_i\right)$. The double stochastically of $D_1 A D_2$ with the extra assumption of uniform marginals means that

$$\forall i = 1,\ldots,n \quad \sum_j d_i^{(1)} d_j^{(2)} a_{ij} = \frac{1}{n}$$

and

$$\forall j = 1,\ldots,n \quad \sum_i d_i^{(1)} d_j^{(2)} a_{ij} = \frac{1}{n}$$

because for any given i and j the selected cell size in the unit square is $1/n$. Hence

$$d_i^{(1)} = \frac{n}{\sum_j d_j^{(2)} a_{ij}}$$

and

$$d_j^{(2)} = \frac{n}{\sum_i d_i^{(1)} a_{ij}}$$

The D_1AD_2 algorithm works by fixed point iteration and is closely related to iterative proportional fitting algorithms.

It can be shown that a multivariate distribution can be arbitrarily well approximated by using a fixed family of bivariate copula. A key step to demonstrating this is to show that the family of bivariate (conditional) copula densities contained in a given multivariate distribution forms a compact set in the space of continuous functions on $[0,1]^2$ (see Bedford et al. (2013) for proof). Based on this it can be shown that the same finite parameter family of copula can be used to give a given level of approximation to all conditional copula simultaneously.

The set $C\left([0,1]^2\right)$ can be considered as a vector space, and in this context a basis is simply sequence of functions $h_1, h_2, \ldots \in C\left([0,1]^2\right)$ for which any function $g \in C\left([0,1]^2\right)$ can be written as $g = \sum_{i=1}^{\infty} \lambda_i h_i$. There are lots of possible bases, for example

$$u, v, uv, u^2, v^2, u^2, vu, v^2, \ldots .$$

Given an ordered basis $h_1, h_2, \ldots \in C\left([0,1]^2\right)$ and a required degree of approximation $\in > 0$ in the sup metric, Bedford et al. (2013) stated the following theorem.

Theorem 2: Given $\in > 0$, there is a k such that any member of $LNC(f)$ can be approximated to within error $\in > 0$ by a linear combination of $h_1, \ldots, h_k$.

First consider a practical guide to build a minimally informative copula structure briefly discussed to approximate any multivariate distribution. A multivariate distribution can be approximated as follows:

- Specify a basic family $B(k)$
- Specify a pair-copula structure
- For each part of pair-copula specify either

1. mean $\alpha_1,...,\alpha_k$ for $h_1,...,h_k$ on each pairwise copula;
2. functions $\alpha_m(ji \mid D_e)$ for the mean values as functions of the conditioning variables, for $m = 1,...,k$, where D_e is the conditioning set for the edge e.

Data set

A data set of Tehran stock exchange is used that includes five time series of daily data: the overall index (*O*), the industry index (*I*), the free float index (*F*), the main board index (*M*) and the secondary index (*S*). All are for the period January 5th2008 to October 30th 2011. (number of observation equal to 668) These five variables are denoted by *O, I, F, M* and *S*, respectively.

First, remove serial correlation in the five time series i.e. the observation of each variable must be independent over time. Hence, the serial correlation in the conditional mean and the conditional variance are modeled by an AR(1) and a GARCH(1,1) model (Bollerslev, 1986), respectively. That is for time series i, the following model for log-return x_i;

$$x_{i,t} = c_i + \alpha_i x_{i,t-1} + \sigma_{i,t} z_{i,t}$$

$$E\left[z_{i,t} \right] = 0$$

and

$$Var\left[z_{i,t} \right] = 1$$

Where $\varepsilon_{i,t-1} = \sigma_{i,t} + z_{i,t}$. Aas et al. (2009)

The further analysis is performed on the standard residuals z_i. If the AR(1)-GARCH(1,1) models are successful at modeling the serial correlation in the conditional mean and the conditional variance, there should be no autocorrelation left in the standard residuals and squared standard residuals. The modified Q-

statistic is used (Ljung and Box, 1979) and the Lagrange Multiplier Test (LM) Engle (1982), respectively, to check this. For all series and both tests, the null hypothesis that there is no autocorrelation left cannot be rejected at the 5% level. Because interest lies mainly in estimating the dependence structure of the risk factor, the standard residual vectors are converted to the uniform variables using the kernel method before further modeling.

It is necessary to generate a vine approximation fitted as in Figure 2 to this data set using minimum information distributions. It should be noted that the corresponding functions of the copula variables *X*, *Y*, *Z*, *U* and *V* associated with *O*, *I*, *F*, *M* and *S* can be found. These are defined by, for example, $h_i(X,Y) = h_i^{'}\left(F_1^{-1}(O), F_2^{-1}(I)\right)$ and should have the same specified expectation, in this case $E[h_i(X,Y)] = E[h_i^{'}(O,I)]$. The minimum information copulas calculated in this example are derived based on copula variables *X*, *Y*, *Z*, *U* and *V*.

It should be noticed that to generalize to other stock exchanges and other applications, a vine structure can be determined uniquely by specifying the order of variables in the first tree T_1. To specify this order, we can use correlation scatter plot, Kendall's τ or the tail dependence coefficient (see e.g., Aas et al., 2009) to measure the strongest bivariate dependencies among the variables in the first tree of the D-vine (or C-vine) of interest. Once the Kendall's τ or the tail dependence coefficients between any pair of the variables in the first tree calculated, then order these measures, and put the variables with the highest measures next to each other and place the ones with weak dependencies farther away. Skipping to present the numerical details of these measures, and following Aas et al. (2009), use the pair-copula construction given in Figure 2 as the selected D-vine structure. In the case, that there is no data to compute these measures to specify the vine structure (or variables order in the first tree), we can use the expert's judgement to elicit these measures or other relevant measures that are more convenient to express by the expert (see Bedford et al. (2013) for a relevant work).

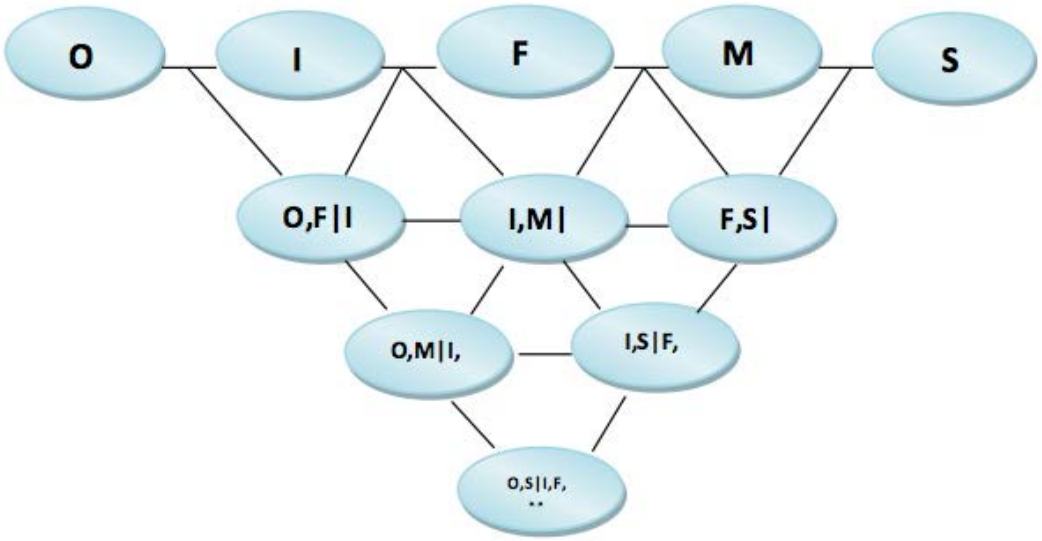


Figure 2. Selected vine structure for the Tehran stock exchange with 5 variables: overall index (O), industry index (I), free float index (F), main board index (M) and secondary index (S).

Initially minimally informative copulas are constructed between each set of two adjacent variables in the first tree, T_1. To do so it is necessary to decide upon which bases to take and how many discretization points to use in each case. The recommended procedure for first copula in T_1, between *O* and *I* is considered next.

Basis function

Which basis functions to include in the copula must first be decided. Basis functions could be chosen, starting with simple polynomials and moving to more complex ones, and including them until satisfied with our approximation. For example if the following basis functions in order is included,

$$OI, O^2I, OI^2, O^2I^2, O^3I, OI^3, O^2I^3, O^3I^2, O^4I, OI^4, OI^5, O^5I, O^3I^3, O^2I^4, O^4I^2$$

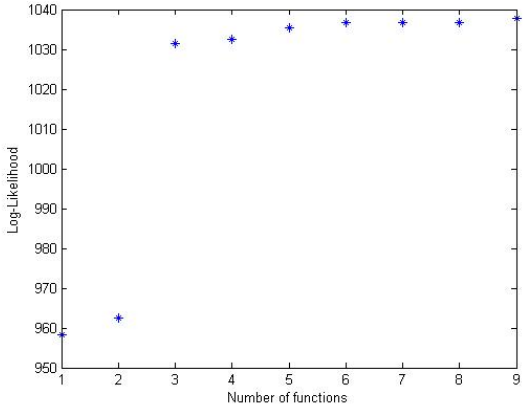


Figure 3. The log-likelihood of the minimally informative copula calculated based on different functions

then the log-likelihood for the copula changes as in Figure 3. There is a jump in the log-likelihood as the third basis function, OI^2 is added. This could imply that we are not adding the basis functions in an optimal manner. Instead at each stage, it is proposed to assess the log-likelihood of adding each additional basis function, then include the function which produces the largest increase in the log-likelihood. Thus the method is similar to a stepwise regression. Doing so for the initial copula yields the basis functions OI, O^2I, O^3I^2.

There is no longer a jump in the log-likelihood when adding the four basis function. The log-likelihood also increase more quickly and reaches its plateau value of around 1030 using fewer basis functions.

Fixing the values of the expectations of these functions by using the empirical data as follows

$$\alpha_1 = \frac{1}{667}\sum_{i=1}^{667} O_i I_i = 0.328,$$

$$\alpha_2 = \frac{1}{667}\sum_{i=1}^{667} O_i^2 I_i = 0.2428,$$

$$\alpha_3 = \frac{1}{667}\sum_{i=1}^{667} O_i^3 I_i^2 = 0.1578$$

The minimum information copula C_{OI} With respect to the uniform distribution given the three constraint above can then be constructed. In order to do so it is necessary to decide on the number of discretization points (or grid size). A larger grid size will provide a better approximation to the continuous copula but at the cost of more computation time. Similarly, the more iteration of the D_1AD_2 and the optimization algorithms that are run, the more accurate the approximation will become. This is again at the expense of speed. Comments on the convergence of the *DAD* algorithm are given in Bedford et al. (2013) and Daneshkhah et al. (2013). In terms of the optimization it is possible to specify how accurate the approximation should be and then judge the effect on the number of iterations required for convergence. In number of iterations needed will also depend on the grid size. In order to be consistent throughout the rest of the example, choose a grid size 50 ◆ 50.

Having done this, the *MI* copula C_{OI} can now be found. This gives parameter value of

$$\lambda_1 = 907.8,$$
$$\lambda_2 = \text{-}1025.1,$$
$$\lambda_3 = 389.41$$

The result has been summarized in table 1 and copula plotted in Figure 4. Note that the Log-likelihood for this copula is 1031.4.

Table 1. Minimum information copula between O and I

Base	Expectation	Parameter Value
OI	0.3280	907.8
O^2I	0.2428	-1025.1
O^3I^2	0.1578	389.41

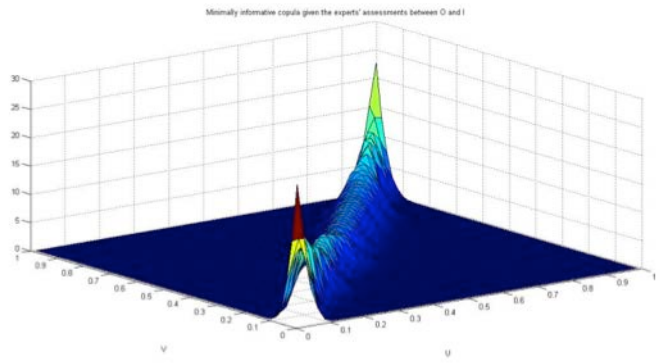


Figure 4. Minimum information copula between *O* and *I*

The second copula in T_1 is C_{IF}. Using the stepwise method as illustrated the following results obtained and the log-likelihood is $l_{IF} = 521.8$. The summarized result are given in Table 2, and Figure 5 shows the fitted copula.

Table 2. Minimum information copula between *I* and *F*.

Base	Expectation	Parameter Value
IF	0.3209	81.2
$I^{3}F^{3}$	0.1254	38.6
$I^{3}F$	0.1851	-75.7

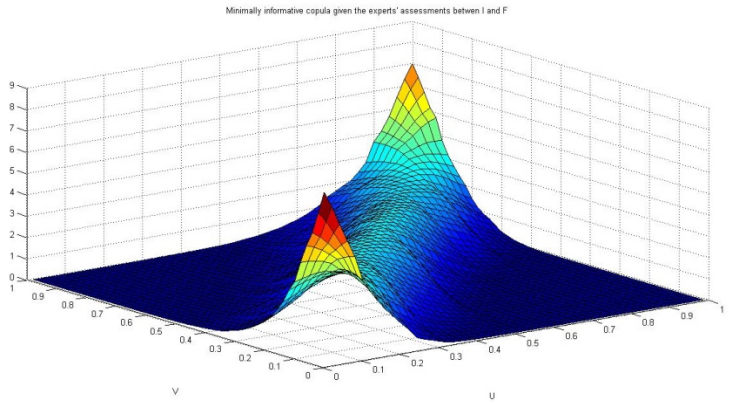


Figure 5. Minimum information copula between *I* and *F*

The third marginal copula is between *F* and *M*. Given a 50×50 grid and a required error of no more than 1×10^{-12} the three bases chosen using the stepwise procedure, the constraint for each base and the resulting parameter values are given in Table 3 and Figure 6 . The log-likelihood for this copula is 462.31.

Table 3. Minimum information copula between *F* and *M*

Base	Expectation	Parameter Value
FM	0.3195	60.97
F^4M^2	0.1252	26.42
FM^3	0.1839	-45.46

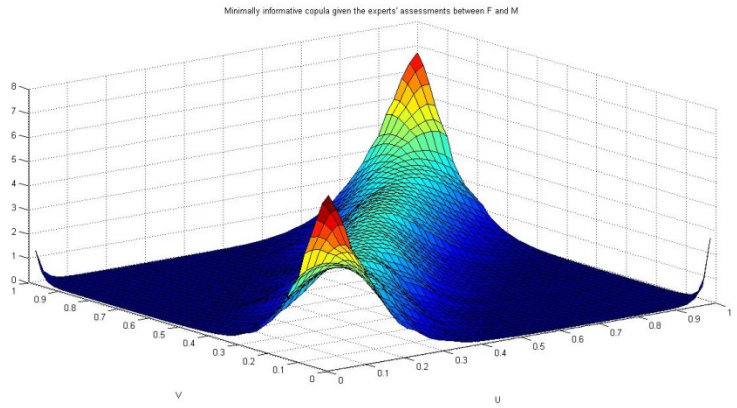


Figure 6. Minimum information copula between *F* and *M*

and the last copula in first tree, T_1, between *M* and *S* is C_{MS} . The result are summarized in Table 4 and Figure 7.

Table 4: Minimum information copula between M and S

Base	Expectation	Parameter Value
MS	0.2928	25.52
MS^2	0.2064	-23.22
M^2S^4	0.0989	8.44

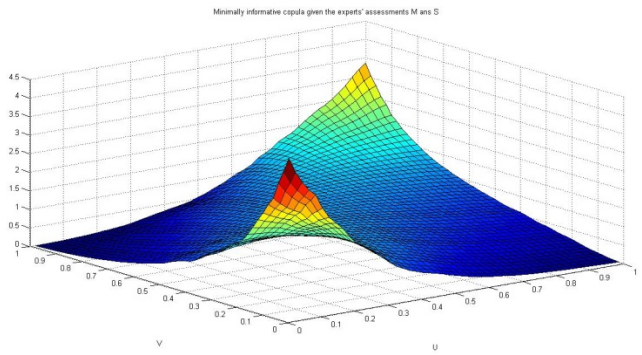


Figure 7: Minimum information copula between *M* and *S*

The conditional copulas in the second tree, T_2, can similarly be approximated using the minimum information approach. Initially the conditional *MI* copula between *O*|*I* and *F*|*I* is constructed. In order to calculate this copula, divide the support of *I* into some arbitrary sub-intervals or bins and then construct the conditional copula within each bin. To do so, find bases in the same way as for the marginal copulas and fit the copulas to the expectations calculated for these. Two bins are used so that the first copula is for *O*,*F*|*I*∈ (0,0.5). The bases for this copula are

$$h_1'\left(O,F \mid I \in (0,0.5)\right) = OF, h_2'\left(O,F \mid I \in (0,0.5)\right) = OF^2,$$
$$h_3' = \left(O,F \mid I \in (0,0.5)\right) = OF^3$$

The expectations given these basis functions which will constrain the *MI* copula are

$$\alpha_1 = 0.0902, \quad \alpha_2 = 0.0368, \quad \alpha_3 = 0.168$$

This process can be followed again for the remaining bins. Table 5 shows the constraints and corresponding Lagrange multipliers required to build the conditional *MI* copula between *O*|*I*∈ (0.5,1) and *F*|*I*∈ (0.5,1). It also gives the log-likelihood in each case.

Table 5. Minimum information copula between *O* and *F* given *I*

Condition	Base	Expectation	Parameter Value	Log-likelihood
0< I <= 0. 5	(OF, OF^2, OF^3)	(.0902,.0368,.0168)	(274.76,-627.3 , 482.6)	195.94
0. 5< I <= 1	(O^2F^4,O^5F, O^4F^2)	(0.247,0.254,0.247)	(-18.2,-69.98 , 81.93)	162.92

Similarly, the *MI* copula can be constructed between remaining nodes in T_2, one of them *I*|*F* and *M*|*F* and another between *F*|*M* and *S*|*M* based on 2 bins and 3 constraints found in the usual manner. The resulting *MI* copula are summarized in Table 6 and Table 7.

Table 6. Minimum information copula between *I* and *M* given *F*

Condition	Base	Expectation	Parameter Value	Log-likelihood
0< F <= 0.5	(IM, IM2, I4M2)	(0.1193,0.06,0.017)	(982.3 , -881.7 , 298.2)	551.3
0.5<F<= 1	(I3M3, I4M2, I4M)	(0.258,0.259,0.302)	(704.4 , -242.1 , -216.8)	555.4

Table 7. Minimum information copula between *F* and *S* given *M*

Condition	Base	Expectation	Parameter Value	Log-likelihood
0<M<= 0.5	(FS, F3S, FS2)	(0.1314,0.0258,0.078)	(51.3 , -51.9 , -11.3)	87
0.5<M<= 1	(F5S, FS5, F3S3)	(0.222,0.197,0.193)	(5.3 , -5.3 , 6.5)	73.7

O|(*I*,*F*) and *M*|(*I*,*F*) are calculated on each combination of bins for *I*,*F*. Thus in T_3 there are 4 bins altogether. The bins, bases and log-likelihoods (l) associated with each copula are given in Table 8.

Similarly the *MI* copulas for *I*|(*F*,*M*) and *S*|(*F*,*M*) are calculated on each combination of bins for *F*,*M*. Table 9 shows the result in this case.

Table 8. Minimum information copula between *O* and *M* given *I* and *F*

Condition	Base	Expectation	Parameter Value	*I*
I <=0.5& F <=0.5	(OM, OM^2, OM^3)	(.082,.031,.0124)	(2685.4 , -7892.9 , 783)	405.95
I <=0.5 & F >0.5	(OM, O^5M, O^4M^2)	(0.164,0.006,0.007)	(2046.3 , -27710 , 1263)	41.9
I >0.5 & F <=0.5	(OM^5, OM^3, O^2M^4)	(0.153,0.245,0.152)	(1481 , -206 , -556)	37.9
I >0.5 & F >0.5	(O^5M, OM, OM^3)	(0.282, 0.582, 0.39)	(728.4, -2054.7, 1025.2)	243.4

Table 9. Minimum information copula between *I* and *S* given *F* and *M*

Condition	Base	Expectation	Parameter Value	*I*
F <=0.5 & M <=0.5	(IS, IS2, IS3)	(.0994,.0542,.0345)	(108.9 , -190.2 , 3243.1)	92.3
F <=0.5 & M >0.5	(I2S, IS2, I5S)	(0.202,0.17,0.061)	(32.2 , -5.6 , -16.5)	6
F >0.5 & M <=0.5	(IS2, I2S3, I2S4)	(0.218,0.082,0.067)	70.9 , -72.9 , 26.7)	4.7
F >0.5 & M >0.5	(I4S2, I3S, I2S)	(0.233, 0.344, 0.42)	(7.5, 2.8,-0.4)	79.9

Table 10. Minimum information copula between O and S given I,F and M

Condition	Base	Expectation	Parameter Value	*I*
I <=0.5&F <=0.5& M <=0.5	(OS, OS2, OS3)	(.0976,.0503,.03)	(111.7 , -173.6 , 80.7)	81.1
I <=0.5 & F <=0.5& M >0.5	(O5S, OS5, OS4)	(0.005,0.002,0.001)	(229.5 , -476.5 , -640)	3.7
I <=0.5 & F >0.5& M <=0.5	(OS, O4S, O4S2)	(0.207,0.014,0.008)	(44.7, -211.8 , 16.6)	1.9
I <=0.5 & F >0.5& M >0.5	(OS5, OS, O4S2)	(0.001, 0.12, 0.003)	(22.9, -985.7, 253.5)	0.1
I >0.5 & F <=0.5& M <=0.5	(O4S, O3S, O2S)	(0.131, 0.203, 0.321)	(17.5, 23.8,- 36.4)	0.2
I >0.5 & F <=0.5& M >0.5	(O3S, OS, O4S2)	(0.194, 0.36, 0.095)	(11.5, -8.5,- 1.1)	0.72
I >0.5 & F >0.5& M <=0.5	(O2S4, O2S2, O2S3)	(0.136, 0.185, 0.158)	(722.7, -211.9,- 633)	0.84
I >0.5 & F >0.5& M >0.5	(O3S3, OS4, OS)	(0.226, 0.266, 0.519)	(11.8, -8.4,-3.14)	77.1

The conditionally *MI* copula in the fourth tree, T_4, can be obtained. In this situation, first divide each of the conditioning variables' supports into 2 bins as in T_2 and T_3, then the *MI* copulas for *O*|(*I*,*F*,*M*) and *S*|(*I*,*F*,*M*) are calculated on each combination of bins for *I*,*F*,*M*. Thus in T_4 there are 8 bins altogether. The bins, bases and log-likelihoods associated with each copula are given in Table 10.

Comparison to the other approaches

Table 11. Comparison to the other approaches

Type of Copula	**Variables (X,Y)**	**Parameters**	***l***
Gaussian copula	(O,I)-(I,F)-(F,M)-(M,S) (O|I,F|I)-(I|F,M|F)-(F|M,S|M) (O|I,F,M|I,F) (I|F,M,S|F,M) (O|I,F,M,S|I,F,M)	Gaussian copula are used as building blocks	3721.04
t-copula	(O,I)-(I,F)-(F,M)-(M,S) (O|I,F|I)-(I|F,M|F)-(F|M,S|M) (O|I,F,M|I,F) (I|F,M,S|F,M) (O|I,F,M,S|I,F,M)	t- copula are used as building blocks	3987.1
MI copula	(O,I)-(I,F)-(F,M)-(M,S) (O|I,F|I)-(I|F,M|F)-(F|M,S|M) (O|I,F,M|I,F) (I|F,M,S|F,M) (O|I,F,M,S|I,F,M)	Details are provided in this article	4845.12

As mentioned, multivariate copula function are limited and weak to modeling multivariate dependency, the proposed method was compared with two different multivariate copula function. When the multivariate Gaussian copula was fit to this data the Log-likelihood is 3458.7 and by multivariate t-copula is 3468.4. In order to make a comparison the log-likelihood of the data sample was computed

for three different copula models used on the same vine structure: The Gaussian copula, the t-copula used by Aas (2009), and our minimum information copula. The results are shown in Table 11.

Conclusion

If choices of marginal densities are made for any indexes of Tehran stock exchange and copulas between them then the above formula will give a multivariate density for each proposed level of variables.

Acknowledgements

The authors are grateful to Professor Tim Bedford, Dr. Kevin Wilson and Dr. Kjersti Aas for helpful comments.

References

Aas, K., Czado, K. C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance, Mathematics and Economics, 44*, 182–198.

Abbas, A. E., Budescu, D. V., & Gu, Y. H. (2010). Assessing joint distributions with isoprobability contours. *Management Science, 56*, 997–1011.

Bedford, T. & Cooke. R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence, 32*, 245-268.

Bedford, T., & Cooke., R. M. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics, 30*(4), 1031–1068.

Bedford, T., & Meeuwissen, A. (1997). Minimally informative distributions with given rank correlation for use in uncertainty analysis. *Journal of Statistical Computation and Simulation, 57*(1- 4), 143 - 174.

Bedford, T., Daneshkhah, A., & Wilson, K. (2013). Approximate Uncertainty Modeling with Vine copulas. To appear in *European Journal of Operation Research.*

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics , 31,* 307–327.

Daneshkhah, A., Parham, G., Chatrabgoun, O., & Jokar, M. (2013). Approximation Multivariate Distribution with pair copula Using the Orthonormal Polynomial and Legendre Multiwavelets basis functions. Submitted to the *Journal of Communications in Statistics - Simulation and Computation.*

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom. *Econometrica, 50*(4), 987–1007.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Kurowicka, D., & Cooke, R. (2006). *Uncertainty analysis with high dimensional dependence modelling*. New York: John Wiley.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303.

Nelsen, R. B. (1999). *An introduction to copulas*. New York: Springer-Verlag, Inc.

Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Universite de Paris, 8*, 229–231.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 427-435.

ISSN 1538 − 9472

Emerging Scholars: On Some Properties of a Heterogeneous Transfer Function Involving Symmetric Saturated Linear (SATLINS) with Hyperbolic Tangent (TANH) Transfer Functions

Christopher Godwin Udomboso
University of Ibadan
Ibadan, Nigeria

For transfer functions to map the input layer of the statistical neural network model to the output layer perfectly, they must lie within bounds that characterize probability distributions. The heterogeneous transfer function, SATLINS_TANH, is established as a Probability Distribution Function (*p.d.f*), and its mean and variance are shown.

Keywords: Statistical neural network, SATLINS, TANH, SATLINS_TANH, mean, variance

Introduction

Anders (1996) proposed a statistical neural network model given as

$$y = f\left(X, w\right) + u \tag{1}$$

where y is the dependent variable, $X = \left(x_0 \equiv 1, x_1, \ldots, x_I\right)$ is a vector of independent variables, $w = (\alpha, \beta, \gamma)$ is the network weight: α is the weight of the input unit, β is the weight of the hidden unit, and γ is the weight of the output unit, and u_i is the stochastic term that is normally distributed (that is, $u_i \sim N(0, \sigma^2 I_n)$).

Basically, $f(X, w)$ is the artificial neural network function, expressed as

$$f\left(X, w\right) = \alpha X + \sum_{h=1}^{H} \beta_h g\left(\sum_{i=0}^{I} \gamma_{hi} x_i\right)$$

The author is a lecturer in the Department of Statistics. Email him at: cg.udomboso@gmail.com.

where $g(.)$ is the transfer function.

The proposed convoluted form of the artificial neural network function used in this study is

$$f\left(\mathrm{X},\mathrm{w}\right)=\alpha\mathrm{X}+\sum_{h=1}^{H}\beta_h\left[g_1\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)g_2\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)\right]$$

and thus, the form of the statistical neural network model proposed is

$$y=\alpha X+\sum_{h=1}^{H}\beta_h\left[g_1\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)g_2\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)\right]+u_iu_j \tag{2}$$

where y is the dependent variable, $X=\left(x_0\equiv1,x_1,\ldots,x_I\right)$ is a vector of independent variables, $w = (\alpha, \beta, \gamma)$ is the network weight: α is the weight of the input unit, β is the weight of the hidden unit, and γ is the weight of the output unit, u_i and u_j are the stochastic terms that are normally distributed (that is, $u_i, u_j \sim N(0, \sigma^2 I_n)$) and $g_1(.)$ and $g_2(.)$ are the transfer functions.

The distributional properties of the heterogeneous model arising from the convolution of SATLINS and TANH is investigated here. Let $g_1(.)$ = Symmetric Saturated Linear function (SATLINS), defined as

$$satlins=g_1\left(.\right)=f_1\left(n\right)=\begin{cases}-1, & n<-1\\ n, & -1\le n\le1\\ 1, & n>1\end{cases} \tag{3}$$

Let $g_2(.)$ = Hyperbolic Tangent function (TANH), defined as

$$\tanh=g_2\left(.\right)=f_2\left(n\right)=\frac{e^n-e^{-n}}{e^n+e^{-n}} \tag{4}$$

Symmetric Saturating Linear and Hyperbolic Tangent

(i)

Let $$f(n) = f_1(n) \otimes f_2(n) = \int_a^b f_1(n-m)\, f_2(m)\, dm \tag{5}$$

For $n < -1$, $f_1(n) = -1$, which implies also that $f_1(n - m) = -1$.

$$f_2(m) = \frac{e^m - e^{-m}}{e^m + e^{-m}}$$

Therefore,

$$\begin{aligned} f_1(n) \otimes f_2(n) &= \int_{-r}^{n} (-1)\left(\frac{e^m - e^{-m}}{e^m + e^{-m}}\right) dm, \qquad r < n < -1 \\ &= \log\left(e^m + e^{-m}\right)^{-1}\Big|_r^n = \log\left(\frac{e^r + e^{-r}}{e^n + e^{-n}}\right) \end{aligned} \tag{6}$$

(ii)

Similarly, for $-1 \le n \le 1$, $f_1(n) = n$, which implies that $f_1(n - m) = n - m$, such that $-1 \le m \le n$.

Therefore,

$$\begin{aligned} f_1(n) \otimes f_1(n) &= \int_{-1}^{n} f_1(n-m)\, f_2(m)\, dm \\ &= \int_{-1}^{n} (n-m)\left(\frac{e^m - e^{-m}}{e^m + e^{-m}}\right) dm \end{aligned} \tag{7}$$

Using integration by part, and noting that

$$\int uv' = uv - \int u'v$$

Let $u = n - m$. This implies that $du = -dm$.

and $v' = \dfrac{d\left[e^{m} + e^{-m}\right]}{e^{m} + e^{-m}}$. This implies that $v = \log\left(e^{m} + e^{-m}\right)$.

Thus,

$$f_1(n) \otimes f_2(n) = (n-m)\log\left(e^{m} + e^{-m}\right) + \int_{-1}^{n} \log\left(e^{m} + e^{-m}\right) dm \tag{8}$$

In (6), let $I = \int_{-1}^{n} \log\left(e^{m} + e^{-m}\right) dm$

Now, let $x = \log\left(e^{m} + e^{-m}\right)$, which implies that $e^{x} = e^{m} + e^{-m}$

But $x = k \in \mathbb{N}$ for $-1 \le m \le 1$. Hence $I = 0$.

Therefore,

$$f_1(n) \otimes f_2(n) = -(n+1)\log\left(e + e^{-1}\right) \tag{9}$$

(iii)

Also, for $n > 1$, $f_1(n) = a = 1$. This implies that $f_1(n - m) = 1$

Therefore,

$$\begin{aligned} f_1(n) \otimes f_2(n) &= \int_{1}^{n} f_1(n-m)\, f_2(m)\, dm \\ &= \int_{1}^{n} \frac{e^{m} - e^{-m}}{e^{m} + e^{-m}} dm = log\left(\frac{e^{n} + e^{-n}}{e + e^{-1}}\right) \end{aligned} \tag{10}$$

The summary of the derived function is given as

$$g_1\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)g_2\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)=f(n)=\begin{cases}\log\left(\dfrac{e^{r}+e^{-r}}{e^{n}+e^{-n}}\right), & \text{for } n<-1\\ (n+1)\log\left(e+e^{-1}\right)^{-1}, & \text{for } -1\le n\le 1\\ \log\left(\dfrac{e^{n}+e^{-n}}{e+e^{-1}}\right), & \text{for } n>1\end{cases} \tag{11}$$

(11) is the derived transfer function for the *Symmetric Saturated Linear transfer function* and the *Hyperbolic Tangent transfer function*.

Distributional Properties of the SATLINS_TANH SNN Model

Next it is shown that the derived transfer functions are probability density functions. By definition, the probability density function (*p.d.f*) of function *f*(*x*) of a random variable $X:\Omega\to\mathbb{R}$ is said to be a proper *p.d.f* if for $x\in(-\infty,+\infty), x\in X$, thus,

$$\int_{-\infty}^{\infty} f(x)\,dx=1,\quad x\in X$$

From the derived transfer function in (11),

$$\int_{-\infty}^{\infty} f_1(n)\otimes f_2(n)\,dn$$
$$=\int_{-\infty}^{-1}\log\left(\frac{e^{r}+e^{-r}}{e^{n}+e^{-n}}\right)dn+\int_{-1}^{1}(n+1)\log\left(e+e^{-1}\right)^{-1}dn$$
$$+\int_{1}^{\infty}\log\left(\frac{e^{n}+e^{-n}}{e+e^{-1}}\right)dn$$

$$= \int_{-\infty}^{-1} \left[\log\left(e^{r} + e^{-r}\right) - \log\left(e^{n} + e^{-n}\right) \right] dn$$

$$+ \log\left(e + e^{-1}\right)^{-1} \int_{-1}^{1} (n+1)\, dn$$

$$+ \int_{1}^{\infty} \left[\log\left(e^{n} + e^{-n}\right) - log\left(e + e^{-1}\right) \right] dn$$

$$= \int_{-\infty}^{-1} \log\left(e^{r} + e^{-r}\right) dn + \log\left(e + e^{-1}\right)^{-1} \int_{-1}^{1} (n+1)\, dn - \int_{1}^{\infty} \log\left(e + e^{-1}\right) dn \qquad (12)$$

$$= \left[n \log\left(e^{r} + e^{-r}\right) \right]_{-\infty}^{-1} + \left[\log\left(e + e^{-1}\right) \left(\frac{n^2}{2} + n \right) \right]_{=1}^{1} - \left[n \log\left(e + e^{-1}\right) \right]_{1}^{\infty}$$

$$= \infty + 2\log\left(e + e^{-1}\right) - \infty$$

$$= 2\log\left(e + e^{-1}\right)$$

The mean and variance of the derived transfer functions are obtained next.

For $f_1(n) \otimes f_2(n)$

$$f_1(n) \otimes f_2(n) = \begin{cases} \log\left(\dfrac{e^{r} + e^{-r}}{e^{n} + e^{-n}}\right), & \text{for } n < -1 \\ (n+1)\log\left(e + e^{-1}\right)^{-1}, & \text{for } -1 \le n \le 1 \\ \log\left(\dfrac{e^{n} + e^{-n}}{e + e^{-1}}\right), & \text{for } n > 1 \end{cases}$$

$$E(n) = \int_{-\infty}^{\infty} n\left(f_1(n) \otimes f_2(n) \right) dn$$

$$= \int_{-\infty}^{-1} n \log\left(\frac{e^{r} + e^{-r}}{e^{n} + e^{-n}}\right) dn + \int_{-1}^{1} n(n+1) \log\left(e + e^{-1}\right)^{-1} dn + \int_{1}^{\infty} n \log\left(\frac{e^{n} + e^{-n}}{e + e^{-1}}\right) dn$$

$$
\begin{aligned}
&= \int_{-\infty}^{-1} n \log\left(e^{r} + e^{-r}\right) dn - \int_{-\infty}^{-1} n \log\left(e^{n} + e^{-n}\right) dn \\
&\qquad + \int_{-1}^{1} n(n+1) \log\left(e + e^{-1}\right)^{-1} dn \\
&\qquad + \int_{1}^{\infty} n \log\left(e^{n} + e^{-n}\right) dn - \int_{1}^{\infty} n \log\left(e + e^{-1}\right) dn \\
&= \log\left(e^{r} + e^{-r}\right) \int_{-\infty}^{-1} n(dn) - \log\left(e + e^{-1}\right)^{-1} \int_{-1}^{1} \left(n^{2} + n\right) dn - \log\left(e + e^{-1}\right) \int_{1}^{\infty} n(dn) \\
&= \log\left(e^{r} + e^{-r}\right) \left[\frac{n^{2}}{2}\right]_{-\infty}^{1} - \log\left(e + e^{-1}\right)^{-1} \left[\frac{n^{3}}{3} + \frac{n^{2}}{2}\right]_{-1}^{1} - \log\left(e + e^{-1}\right) \left[\frac{n^{2}}{2}\right]_{1}^{\infty} \\
&= \log\left(e^{r} + e^{-r}\right) \left(\frac{1}{2} - \frac{\infty}{2}\right) - \log\left(e + e^{-1}\right)^{-1} \left[\left(\frac{1}{3} + \frac{1}{2}\right) - \left(\frac{1}{2} - \frac{1}{3}\right)\right] - \log\left(e + e^{-1}\right) \left(\frac{\infty}{2} - \frac{1}{2}\right)
\end{aligned}
$$

Hence, the mean of derived transfer function in is

$$E(n) = \frac{2}{3} \log\left(e + e^{-1}\right)^{-1} \tag{13}$$

Similarly,

$$
\begin{aligned}
E\left(n^{2}\right) &= \int_{-\infty}^{\infty} n^{2} \left(f_{1}(n) \otimes f_{2}(n)\right) dn \\
&= \int_{-\infty}^{-1} n^{2} \log\left(\frac{e^{r} + e^{-r}}{e^{n} + e^{-n}}\right) dn + \int_{-1}^{1} n^{2} (n+1) \log\left(e + e^{-1}\right)^{-1} dn + \int_{1}^{\infty} n^{2} \log\left(\frac{e^{n} + e^{-n}}{e + e^{-1}}\right) dn \\
&= \log\left(e^{r} + e^{-r}\right) \int_{-\infty}^{-1} n^{2} dn - \int_{-\infty}^{-1} n^{2} \log\left(e^{n} + e^{-n}\right) dn \\
&\qquad + \log\left(e + e^{-1}\right)^{-1} \int_{-1}^{1} \left(n^{3} - n^{2}\right) dn \\
&\qquad + \int_{1}^{\infty} n^{2} \log\left(e^{n} + e^{-n}\right) dn - \log\left(e + e^{-1}\right) \int_{1}^{\infty} n^{2} dn
\end{aligned}
$$

$$= \log\left(e^{r} + e^{-r}\right)\left[\frac{n^3}{3}\right]_{-\infty}^{-1} + \log\left(e + e^{-1}\right)^{-1}\left[\frac{n^4}{4} - \frac{n^3}{3}\right]_{-1}^{1} + \log\left(e + e^{-1}\right)\left[\frac{n^3}{3}\right]_{1}^{\infty}$$

$$= \log\left(e^{r} + e^{-r}\right)\left(\frac{\infty}{3} - \frac{1}{3}\right) + \log\left(e + e^{-1}\right)^{-1}\left[\left(\frac{1}{4} + \frac{1}{3}\right) - \left(\frac{1}{4} - \frac{1}{3}\right)\right] - \log\left(e + e^{-1}\right)\left(\frac{\infty}{3} - \frac{1}{3}\right)$$

$$= \frac{2}{3}\log\left(e + e^{-1}\right)^{-1}$$

Therefore, variance of $\left(f_1(n) \otimes f_2(n)\right)$ is

$$\begin{aligned} \operatorname{var}(n) &= E\left(n^2\right) - \left[E(n)\right]^2 \\ &= \frac{2}{3}\log\left(e + e^{-1}\right)^{-1} - \left[\frac{2}{3}\log\left(e + e^{-1}\right)^{-1}\right]^2 \\ &= -\frac{2}{3}\log\left(e + e^{-1}\right) + \frac{4}{9}\left(\log\left(e + e^{-1}\right)\right)^2 \\ &= \log\left(e + e^{-1}\right)\left[\frac{4}{9}\log\left(e + e^{-1}\right) - \frac{2}{3}\right] \end{aligned} \tag{14}$$

Thus,

$$g_1\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right)g_2\left(\sum_{i=0}^{I}\gamma_{hi}x_i\right) = f(n) = \begin{cases} \log\left(\dfrac{e^{r} + e^{-r}}{e^{n} + e^{-n}}\right), & \text{for } n < -1 \\ (n+1)\log\left(e + e^{-1}\right)^{-1}, & \text{for } -1 \le n \le 1 \\ \log\left(\dfrac{e^{n} + e^{-n}}{e + e^{-1}}\right), & \text{for } n > 1 \end{cases}$$

with mean, $E(n) = \frac{2}{3}\log\left(e + e^{-1}\right)^{-1}$

and variance, $\operatorname{var}(n) = \log\left(e + e^{-1}\right)\left[\frac{4}{9}\log\left(e + e^{-1}\right) - \frac{2}{3}\right]$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Anders, U. (1996). *Statistical Model Building for Neural Networks.* AFIR Colloqium. Nunberg, Germany.

Anderson, J. A. (2003). *An Introduction to neural networks.* Upper Saddle River, NJ: Prentice Hall.

Battiti, R. (1992). First- and second-order methods for learning: Between steepest descent and Newton's method. *Neural Computation 4*, 141–166.

Carling, A. (1992). *Introducing Neural Networks*. Ammanford, England: Sigma Press.

Foresee F. D., & Hagan, M. T. (1997). 'Gauss-Newton Approximation to Bayesian Regularization.' In *IEEE International Conference on Neural Networks (Vol. 3)*, 1930–1935. New York: IEEE.

Golden, R.M. (1996). *Mathematical Methods for Neural Network Analysis and Design.* Cambridge, MA: MIT Press.

Lawrence, J. (1994). *Introduction to neural networks: Design, theory, and applications.* Nevada City, CA: California Scientific Software Press.

Maren, A., Harston, C., & Pap, R., (1990). *Handbook of Neural Computing Applications*. San Diego, CA: Academic Press.

Nelson, M. M. & Illingworth, W. T. (1991). *A Practical Guide to Neural Nets*. Addison-Wesley Publishing Company.

Smith, M. (1993). *Neural networks for statistical modeling*. New York: Van Nostrand Reinhold.

Taylor, J. G. (1999). *Neural networks and their applications*. New York: Wiley.

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The American Statistician, 50*(4), 284–293.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 436-449.

ISSN 1538 − 9472

Emerging Scholars: Distribution of the Ratio of Normal and Rice Random Variables

Nayereh B. Khoolenjani
University of Isfahan
Isfahan, Iran

Kavoos Khorshidian
Shiraz University
Shiraz, Iran

The ratio of independent random variables arises in many applied problems. The distribution of the ratio $\left|X/Y\right|$ is studied when X and Y are independent Normal and Rice random variables, respectively. Ratios of such random variables have extensive applications in the analysis of noises in communication systems. The exact forms of probability density function (PDF), cumulative distribution function (CDF) and the existing moments are derived in terms of several special functions. As a special case, the PDF and CDF of the ratio of independent standard Normal and Rayleigh random variables have been obtained. Tabulations of associated percentage points and a computer program for generating tabulations are also given.

Keywords: Normal distribution, Rice distribution, ratio random variable, special functions.

Introduction

For given random variables X and Y, the distribution of the ratio $\left|X/Y\right|$ arises in a wide range of natural phenomena of interest, such as in engineering, hydrology, medicine, number theory, psychology, etc. More specifically, Mendelian inheritance ratios in genetics, mass to energy ratios in nuclear physics, target to control precipitation in meteorology, inventory ratios in economics are exactly of this type. The distribution of the ratio random variables (RRV) has been extensively investigated by many authors especially when X and Y are independent and belong to the same family. Various methods have been compared and reviewed by authors including Pearson (1910), Greay (1930), Marsaglia (1965, 2006) and Nadarajah (2006).

Nayereh B. Khoolenjani is a Ph.D. student in the Department of Statistics. Email at: n.b.khoolenjani@gmail.com. Kavoos Khorshidian is in the Department of Statistics.

The exact distribution of $\left|X/Y\right|$ is derived when X and Y are independent random variables (RVs) having Normal and Rice distributions with parameters (μ,σ^2) and (λ,ν), respectively. The Normal and Rice distributions are well known and of common use in engineering, especially in signal processing and communication theory. In engineering, there are many real situations where measurements could be modeled by Normal and Rice distributions. Some typical situations in which the ratio of Normal and Rice random variables appear are as follows. In the case that X and Y represent the random noises corresponding to two signals, studying the distribution of the quotient $\left|X/Y\right|$ is of interest. For example in communication theory it may represent the relative strength of two different signals and in MRI, it may represent the quality of images. Moreover, because of the important concept of moments of RVs as magnitude of power and energy in physical and engineering sciences, the possible moments of the ratio of Normal and Rice random variables have been also obtained. Applications of Normal and Rice distributions and the ratio RVs may be found in Rice (1974), Helstrom (1997), Karagiannidis and Kotsopoulos (2001), Salo, et al. (2006), Withers and Nadarajah (2008) and references therein.

The probability density function (PDF) of a two-parameter Normal random variable X can be written as:

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\},\ -\infty<x<\infty \tag{1}$$

where $-\infty<\mu<\infty$ is the location parameter and $\sigma>0$ is the scale parameter. For $\mu=0$ and $\sigma^2=1$, (1) becomes the distribution of standard Normal random variable. A well known representation for CDF of X is as

$$F_X(x)=\frac{1}{2}\left\{1+erf\,(\frac{x-\mu}{\sigma\sqrt{2}})\right\} \tag{2}$$

where $erf\,(\cdot)$ denotes the error function that is given by

$$erf\,(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}du \tag{3}$$

Also,

$$E(X^k) = \mu^k \cdot k! \sum_{j=0}^{\left[\frac{k}{2}\right]} \frac{1}{(k-2j)!\, j!} \left(\frac{\sigma^2}{2\mu^2}\right)^j. \tag{4}$$

If Y has a Rice distribution with parameters (λ, v) , then the PDF of Y is as follows:

$$f_Y(y) = \frac{y}{\lambda^2} \exp\{-\frac{(y^2+v^2)}{2\lambda^2}\} \mathrm{I}_0\left(\frac{yv}{\lambda^2}\right),\ y > 0,\ v \geq 0,\ \lambda > 0 \tag{5}$$

where y is the signal amplitude, $\mathrm{I}_0(.)$ is the modified Bessel function of the first kind of order 0, $2\lambda^2$ is the average fading-scatter component and v^2 is the line-of-sight (LOS) power component. The Local Mean Power is defined as $\Omega = 2\lambda^2 + v^2$ which equals $E(X^2)$, and the Rice factor K of the envelope is defined as the ratio of the signal power to the scattered power, i.e., $K = v^2 / 2\lambda^2$. When K goes to zero, the channel statistic follows Rayleigh distribution, whereas if K goes to infinity, the channel becomes a non-fading channel. For $v = 0$, the expression (5) reduces to a Rayleigh distribution.

Notations and Preliminaries

Recall some special mathematical functions, these will be used repeatedly throughout this study. The modified Bessel function of first kind of order v , is

$$I_v(x) = \left(\frac{1}{2}x\right)^v \sum_{k=0}^{\infty} \frac{\left(\frac{1}{4}x^2\right)^k}{(k!)\Gamma(v+k+1)} \tag{6}$$

The generalized hypergeometric function is denoted by

$${}_pF_q(a_1, a_2, ..., a_p; b_1, b_2, ..., b_q; z) = \sum_{k=0}^{\infty} \frac{(a_1)_k (a_2)_k ... (a_p)_k}{(b_1)_k (b_2)_k ... (b_q)_k} \frac{z^k}{k!} \tag{7}$$

The Gauss hypergeometric function and the Kummer confluent hypergeometric function are given, respectively, by

$$ {}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!} \tag{8} $$

and

$$ {}_1F_1(a;b;z) = \sum_{k=0}^{\infty} \frac{(a)_k z^k}{(b)_k k!} \tag{9} $$

where $(a)_k$, $(b)_k$ represent Pochhammer's symbol given by

$(a)_k = a(a+1)\cdots(a+k-1) = \dfrac{\Gamma(\alpha+k)}{\Gamma(\alpha)}$.

The parabolic cylinder function is

$$ D_v(z) = 2^{\frac{v}{2}} e^{-\frac{z^2}{4}} \Psi(-\frac{1}{2}v, \frac{1}{2}; \frac{1}{2}z^2) \tag{10} $$

where $\Psi(a,c;z)$ represents the confluent hypergeometric function given by

$$ \Psi(a,c;z) = \Gamma\begin{bmatrix} 1-c \\ 1+a-c \end{bmatrix} {}_1F_1(a;c;z) + \Gamma\begin{bmatrix} c-1 \\ a \end{bmatrix} 2^{1-c} {}_1F_1(1+a-c;2-c;z), $$

in which

$$ \Gamma\begin{bmatrix} a_1,\ldots,a_m \\ b_1,\ldots,b_n \end{bmatrix} = \frac{\prod_{i=1}^{m} \Gamma(a_i)}{\prod_{j=1}^{n} \Gamma(b_j)}. $$

The complementary error function is denoted by

$$erfc(x)=\frac{2}{\sqrt{\pi}}\int_{x}^{\infty}e^{-u^2}du \tag{11}$$

The following lemmas are of frequent use.

Lemma 1 (Equation (2.15.5.4), Prudnikov, et al., 1986). For $\mathrm{Re}\,p>0$, $\mathrm{Re}(\alpha+v)>0;\ |\arg c|<\pi$

$$\int_0^{\infty}x^{\alpha-1}e^{-px^2}I_v(cx)dx$$
$$=2^{-v-1}c^v p^{-\frac{(\alpha+v)}{2}}\Gamma\begin{bmatrix}\frac{(\alpha+v)}{2}\\ v+1\end{bmatrix}{}_1F_1(\frac{\alpha+v}{2};v+1;\frac{c^2}{4p}).$$

Lemma 2 (Equation (2.8.9.2), Prudnikov, et al., 1986). For $\mathrm{Re}\,p>0;|\arg c|<\frac{\pi}{4}$

$$\int_0^{\infty}x^{2n+1}e^{-px^2}\begin{Bmatrix}erf\,(cx+b)\\ erfc(cx+b\end{Bmatrix}dx=\begin{Bmatrix}0\\1\end{Bmatrix}\frac{n!}{2p^{n+1}}\pm\frac{(-1)^n}{2}$$
$$\frac{\partial^n}{\partial P^n}\left[\frac{1}{p}erf\,(b)+\frac{c}{p\sqrt{c^2+p}}\exp(-\frac{pb^2}{c^2+p})erfc(\frac{bc}{\sqrt{c^2+p}})\right].$$

Lemma 3 (Equation (3.462.1), Gradshteyn & Ryzhik, 2000). For $\mathrm{Re}\,\beta>0$, $\mathrm{Re}\,v>0$

$$\int_0^{\infty}x^{v-1}\exp(\left\{-\beta x^2-\gamma x\right\}dx=(2\beta)^{-\frac{v}{2}}\Gamma(v)\exp(\frac{\gamma^2}{8\beta})D_{-v}(\frac{\gamma}{\sqrt{2\beta}}).$$

The Ratio of Normal and Rice Random Variables

The explicit expressions for the PDF and CDF of $|X/Y|$ are derived in terms of the Gauss hypergeometric function. The ratio of standard Normal and Rayleigh RVs is also considered as a special case.

Theorem 1: Suppose that X and Y are independent Normal and Rice random variables with parameters (μ,σ^2) and (λ,ν), respectively. The PDF of the ratio random variable $T=\left|X/Y\right|$ can be expressed as $f(t)=g(t)+g(-t)$, where

$$g(t)=\frac{e^{-\{\frac{\nu^2}{2\lambda^2}+\frac{\mu^2}{2\sigma^2}-\frac{\mu^2t^2\lambda^2}{4\sigma^2(\lambda^2t^2+\sigma^2)}\}}\sigma^2\lambda}{\sqrt{2\pi}(\lambda^2t^2+\sigma^2)^{\frac{3}{2}}} \times\sum_{k=0}^{\infty}\frac{(\frac{\nu^2}{4\lambda^2})^k}{(k!)^2}\cdot\Gamma(2k+3)\cdot D_{-(2k+3)}(\frac{-\mu t\lambda}{\sigma\sqrt{\lambda^2t^2+\sigma^2}}). \tag{12}$$

Theorem 1 Proof:

$$\begin{aligned} f(t)&=\int_0^\infty yf_X(ty)f_Y(y)dy+\int_0^\infty yf_X(-ty)f_Y(y)dy \\ &=\int_0^\infty y\cdot\frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{1}{2\sigma^2}(ty-\mu)^2\right\}\cdot\frac{y}{\lambda^2}exp\left\{-\frac{(y^2+\nu^2)}{2\lambda^2}\right\}I_0(\frac{y\nu}{\lambda^2})dy \\ &+\int_0^\infty y\cdot\frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{1}{2\sigma^2}(-ty-\mu)^2\right\}\cdot\frac{y}{\lambda^2}exp\left\{-\frac{(y^2+\nu^2)}{2\lambda^2}\right\}I_0(\frac{y\nu}{\lambda^2})dy \end{aligned} \tag{13}$$

The two integrals in (13) can be calculated by direct application of Lemma 3. Thus the result follows.

Remark 2: By using expression (10), elementary forms for $g(t)$ in Theorem 1 can be derived as follows:

$$g(t)=\frac{e^{-\frac{1}{2\sigma^2\lambda^2}(\nu^2\sigma^2+\mu^2\lambda^2)}\lambda\sigma^2}{\sqrt{2\pi}(t^2\lambda^2+\sigma^2)^{\frac{3}{2}}}\sum_{k=0}^{\infty}\frac{(\frac{\nu^2}{4\lambda^2})^k\Gamma(2k+3)}{(k!)^2 2^{\frac{2k+3}{2}}}\Psi(\frac{2k+3}{2},\frac{1}{2};\frac{\mu^2t^2\lambda^2}{2\sigma^2(t^2\lambda^2+\sigma^2)}) \quad (14)$$

Corollary 3 Assume that X and Y are independent standard Normal and Rayleigh random variables, respectively. The PDF of the ratio random variable $T=\left|X/Y\right|$ can be expressed as

$$f_T(t)=\frac{\lambda}{(t^2\lambda^2+1)^{3/2}} \quad , t>0 \quad (15)$$

Theorem 4: Suppose that X and Y are independent Normal and Rice random variables with parameters (μ,σ^2) and (λ,ν), respectively. The CDF of the ratio random variable $T=\left|X/Y\right|$ can be expressed as $F(t)=G(t)-G(-t)$ where

$$G(t)=\frac{e^{-\frac{\nu^2}{2\lambda^2}}}{2\lambda^2}\sum_{k=0}^{\infty}\frac{(\frac{\nu^2}{4\lambda^4})^k}{(k!)^2}\{\frac{n!}{2(\frac{1}{2\lambda^2})^{k+1}}-\frac{(-1)^k}{2}\frac{\partial^k}{\partial(\frac{1}{2\lambda^2})^k}[2\lambda^2 erf(\frac{-\mu}{\sqrt{2}\sigma}) \quad (16)$$
$$-\frac{2t\lambda^3}{\sqrt{t^2\lambda^2+\sigma^2}}\times exp(-\frac{\mu^2}{2(t^2\lambda^2+\sigma^2)})erfc(-\frac{\mu t\lambda}{\sigma\sqrt{2(t^2\lambda^2+\sigma^2)}})]\}.$$

Theorem 4 Proof: The CDF $F(t)=\Pr(\left|X/Y\right|\le t)$ can be written as

$$F(t)=\int_0^{\infty}\left\{\Phi(\frac{ty-\mu}{\sigma})-\Phi(\frac{-ty-\mu}{\sigma})\right\}f_Y(y)dy, \quad (17)$$

where $\Phi(.)$ is the cdf of the standard Normal distribution. Using the relationship

$$\Phi(-x)=\frac{1}{2}erfc(\frac{x}{\sqrt{2}}), \quad (18)$$

Eq. (17) can be rewritten as

$$F(t)=\frac{1}{2}\int_0^{\infty}\left\{erfc(\frac{\mu-ty}{\sigma\sqrt{2}})-erfc(\frac{\mu+ty}{\sigma\sqrt{2}})\right\}f_Y(y)\,dy$$
$$=\frac{1}{2}\int_0^{\infty}erfc(\frac{\mu-ty}{\sigma\sqrt{2}})\cdot\frac{y}{\lambda^2}\exp\left\{-\frac{(y^2+v^2)}{2\lambda^2}\right\}I_0(\frac{yv}{\lambda^2})dy \qquad (19)$$
$$-\frac{1}{2}\int_0^{\infty}erfc(\frac{\mu+ty}{\sigma\sqrt{2}})\cdot\frac{y}{\lambda^2}\exp\left\{-\frac{(y^2+v^2)}{2\lambda^2}\right\}I_0(\frac{yv}{\lambda^2})dy.$$

The result follows by using Lemma 2.

Corollary 5: Assume that X and Y are independent Normal and Rice random variables with parameters $(0,\sigma^2)$ and $(\lambda,0)$, respectively. The CDF of the ratio random variable $T=\left|X/Y\right|$ is

$$F(t)=\frac{t\lambda}{\sqrt{t^2\lambda^2+\sigma^2}}\quad,\quad t>0. \qquad (20)$$

Figures (1) and (2) illustrate possible shapes of the pdf corresponding to (20) for different values of σ^2 and λ. Note that the shape of the distribution is mainly controlled by the values of σ^2 and λ.

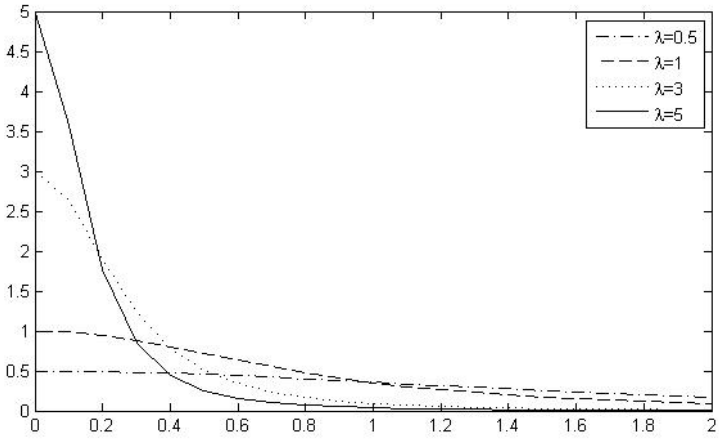


Figure 1 Plots of the pdf corresponding to (20) for $\lambda=0.5,1,3,5$ and $\sigma=1$.

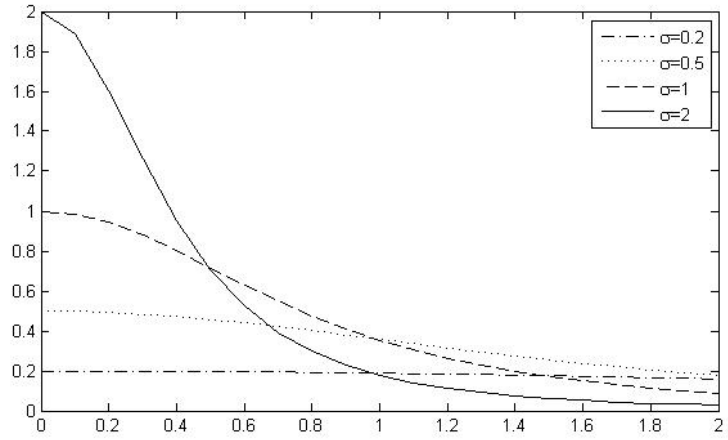


Figure 2 Plots of the pdf corresponding to (20) for $\sigma = 0.2, 0.5, 1, 2$ and $\lambda = 1$.

K^{th} Moments of the Ratio Random Variable

In the sequel, the independence of X and Y are used several times for computing the moments of the ratio random variable. The results obtained are expressed in terms of confluent hypergeometric functions.

Theorem 6: Suppose that X and Y are independent Normal and Rice random variables with parameters (μ, σ^2) and (λ, ν), respectively. A representation for the kth moment of the ratio random variable $T = X/Y$, for $k < 2$, is:

$$E[T^k] = \left(\frac{\mu}{\sqrt{2}\lambda}\right)^k k! e^{-\frac{\nu^2}{2\lambda^2}} \Gamma\left(\frac{-k+2}{2}\right) {}_1F_1\left(\frac{-k+2}{2}; 1; \frac{\nu^2}{2\lambda^2}\right) \sum_{j=0}^{[\frac{k}{2}]} \frac{1}{(k-2j)!\, j!} \left(\frac{\sigma^2}{2\mu^2}\right)^j \quad (21)$$

Theorem 6 Proof: Using the independency of X and Y the expected ratio can be written as

$$E(T^k) = E\left(\frac{X^k}{Y^k}\right) = E(X^k) E\left(\frac{1}{Y^k}\right), \quad (22)$$

in which

$$E(\frac{1}{Y^k}) = \int_0^\infty \frac{1}{y^k} \cdot \frac{y}{\lambda^2} \exp\left\{-\frac{(y^2+v^2)}{2\lambda^2}\right\} I_0(\frac{yv}{\lambda^2}) dy \tag{23}$$

By using lemma 2.1, the integral (23) reduces to

$$E(\frac{1}{Y^k}) = \frac{e^{-\frac{v^2}{2\lambda^2}}}{(2\lambda^2)^{\frac{k}{2}}} \Gamma(\frac{-k+2}{2})\, {}_1F_1(\frac{-k+2}{2}; 1; \frac{v^2}{2\lambda^2}) \tag{24}$$

The desired result now follows by multiplying (4) and (24).

Remark 7: Formula (21), displays the exact forms for calculating $E(T)$, which have been expressed in terms of confluent hypergeometric functions. The delta-method can be used to approximate the first and second moments of the ratio $T = X/Y$. In detail, by taking $\mu_X = E(X)$, $\mu_Y = E(Y)$ and using the Delta-method (Casella & Berger, 2002) results in:

$$E(T) \approx \frac{\mu_X}{\mu_Y} = \sqrt{\frac{2}{\pi}} \frac{\mu e^{\frac{v^2}{2\lambda^2}}}{\lambda\, {}_1F_1(\frac{3}{2}, 1; \frac{v^2}{2\lambda^2})}.$$

For approximating $Var(T)$, first recall that $E(X^2) = \mu^2 + \sigma^2$ and $E(Y^2) = 2\lambda^2 + v^2$. Thus,

$$Var(\frac{X}{Y}) \approx \left(\frac{{\mu_X}^2}{{\mu_Y}^2}\right)\left(\frac{Var(X)}{{\mu_X}^2} + \frac{Var(Y)}{{\mu_Y}^2}\right),$$

which involves confluent hypergeometric functions, but in simpler forms.

Remark 8: The numerical computation of the obtained results in this article entails calculation of special functions, their sums and integrals, which have been tabulated and available in determinds books and computer algebra packages (see Greay, 1930; Helstrom, 1997; and Salo, et al. 2006 for more details.

Percentiles

Table 1. Percentage points of $T = \left| X \middle/ Y \right|$ for $\lambda = 0.1 - 2.5$.

λ	$p = 0.01$	$p = 0.05$	$p = 0.1$	$p = 0.9$	$p = 0.95$	$p = 0.99$
0.1	0.100005	0.500626	1.005023	20.64741	30.4243	70.1792
0.2	0.050002	0.250313	0.502518	10.32370	15.2121	35.0896
0.3	0.033335	0.166875	0.335012	6.882471	10.1414	23.3930
0.4	0.025001	0.125156	0.251259	5.16185	7.6060	17.5448
0.5	0.020001	0.100125	0.201007	4.12948	6.0848	14.0358
0.6	0.016667	0.083437	0.167506	3.44123	5.0707	11.6965
0.7	0.014286	0.071518	0.143576	2.94963	4.3463	10.0256
0.8	0.012503	0.062578	0.125629	2.58092	3.8030	8.7724
0.9	0.011111	0.055625	0.111670	2.29415	3.3804	7.7976
1	0.010002	0.050062	0.100503	2.06474	3.0424	7.0179
1.1	0.009091	0.045511	0.091367	1.87703	2.7658	6.3799
1.2	0.008333	0.041718	0.083753	1.72061	2.5353	5.8482
1.3	0.0076926	0.038509	0.077310	1.58826	2.3403	5.3984
1.4	0.0071432	0.035759	0.071788	1.47481	2.1731	5.0128
1.5	0.0066670	0.033375	0.067002	1.37649	2.0282	4.6786
1.6	0.0062503	0.031289	0.062814	1.29046	1.9015	4.3862
1.7	0.0058826	0.029448	0.059119	1.21455	1.7896	4.1281
1.8	0.0055558	0.027812	0.055835	1.14707	1.6902	3.8988
1.9	0.0052634	0.026348	0.052896	1.08670	1.6012	3.6936
2	0.0050002	0.025031	0.050251	1.03237	1.5212	3.5089
2.1	0.0047621	0.023839	0.047858	0.98321	1.4487	3.3418
2.2	0.0045456	0.022755	0.045683	0.93851	1.3829	3.1899
2.3	0.0043480	0.021766	0.043697	0.89771	1.3227	3.0512
2.4	0.0041668	0.020859	0.041876	0.86030	1.2676	2.9241
2.5	0.0040002	0.020025	0.040201	0.82589	1.2169	2.8071

Table 2. Percentage points of $T = \left| X \middle/ Y \right|$ for $\lambda = 2.6 - 5$.

λ	$p = 0.01$	$p = 0.05$	$p = 0.1$	$p = 0.9$	$p = 0.95$	$p = 0.99$
2.6	0.0038463	0.019254	0.038655	0.79413	1.1701	2.6992
2.7	0.0037038	0.018541	0.037223	0.76471	1.1268	2.5992
2.8	0.0035716	0.017879	0.035894	0.73740	1.0865	2.5064
2.9	0.0034484	0.017262	0.034656	0.71197	1.0491	2.4199
3	0.0033335	0.016687	0.033501	0.68824	1.0141	2.3393
3.1	0.0032259	0.016149	0.032420	0.66604	0.9814	2.2638
3.2	0.0031251	0.015644	0.031407	0.64523	0.9507	2.1931
3.3	0.0030304	0.015170	0.030455	0.62567	0.9219	2.1266
3.4	0.0029413	0.014724	0.029559	0.60727	0.8948	2.0640
3.5	0.0028572	0.014303	0.028715	0.58992	0.8692	2.0051
3.6	0.0027779	0.013906	0.027917	0.57353	0.8451	1.9494
3.7	0.0027028	0.013530	0.027163	0.55803	0.8222	1.8967
3.8	0.0026317	0.013174	0.026448	0.54335	0.8006	1.8468
3.9	0.0025642	0.012836	0.025770	0.52942	0.7801	1.7994
4	0.0025001	0.012515	0.025125	0.51618	0.7606	1.7544
4.1	0.0024391	0.012210	0.024513	0.50359	0.7420	1.7116
4.2	0.0023810	0.011919	0.023929	0.4916	0.7243	1.6709
4.3	0.0023256	0.011642	0.023372	0.48017	0.7075	1.6320
4.4	0.0022728	0.011377	0.022841	0.46925	0.6914	1.5949
4.5	0.0022223	0.011125	0.022334	0.45883	0.6760	1.5595
4.6	0.0021740	0.010883	0.021848	0.44885	0.6613	1.5256
4.7	0.0021277	0.010651	0.021383	0.43930	0.6473	1.4931
4.8	0.0021145	0.010532	0.020672	0.42654	0.6311	1.4752
4.9	0.0021073	0.010380	0.019823	0.41839	0.6277	1.4613
5	0.0020094	0.010157	0.018782	0.41027	0.6120	1.4479

Tabulations of percentage points t_p associated with the cdf (20) of $T = \left| X \middle/ Y \right|$ are provided. These values are obtained by numerically solving:

$$\frac{t_p \lambda}{\sqrt{t_p{}^2 \lambda^2 + \sigma^2}} = p \tag{25}$$

Tables 1 and 2 provide the numerical values of t_p for $\lambda = 0.1, 0.2, ..., 5$ and $\sigma = 1$. It is hoped that these numbers will be of use to practitioners as mentioned in the introduction. Similar tabulations could be easily derived for other values of λ, σ and p by using the sample program provided in Appendix A.

References

Casella, G., & Berger, L. B. (2002), *Statistical Inference*. Duxbury Press.

Gradshteyn, I. S., & Ryzhik, I. M. (2000).*Table of Integrals, Series, and Products.* San Diego, CA: Academic Press.

Greay R. C. (1930). The frequency distribution of the quotient of two normal variates. *Journal of the Royal Statistical Society. 93*, 442-446.

Helstrom, C. (1997). Computing the distribution of sums of random sine waves and the Rayleigh-distributed random variables by saddle-point integration. *IEEE Trans. Commun. 45*(11), 1487-1494.

Karagiannidis, G. K., & Kotsopoulos, S. A. (2001). On the distribution of the weighted sum of L independent Rician and Nakagami Envelopes in the presence of AWGN. *Journal of Communication and Networks. 3*(2), 112-119.

Marsaglia, G. (1965). Ratios of Normal Variables and Ratios of Sums of Uniform. *JASA. 60*, 193, 204.

Marsaglia, G. (2006). Ratios of Normal Variables. *Journal of Statistical Software. 16*(4), 1-10.

Nadarajah, S. (2006). Quotient of Laplace and Gumbel random variables. *Mathematical Problems in Engineering.* vol. 2006. Article ID 90598, 7 pages.

Pearson, K. (1910). On the constants of Index-Distributions as Deduced from the Like Constants for the Components of the Ratio with Special Reference to the Opsonic Index. *Biometrika. 7*(4), 531-541.

Prudnikov, A. P., Brychkov Y.A., Marichev O.I. (1986). *Integrals and Series, 2.* New York: Gordon and Breach.

Rice, S. (1974). Probability distributions for noise plus several sin waves the problem of computation. *IEEE Trans. Commun.* 851-853.

Salo, J., EI-Sallabi, H. M., & Vainikainen, P. (2006). The distribution of the product of independent Rayleigh random variables. *IEEE Transactions on Antennas and Propagation. 54*(2), 639-943.

Withers, C. S. & Nadarajah, S. (2008). MGFs for Rayleigh Random Variables. *Wireless Personal Communications. 46*(4), 463-468.

Appendix A

The following program in R can be used to generate tables similar to that presented in the section headed 'Percentiles.'

```
p=c(0.01,.05,0.1,0.9,0.95,0.99)
sig=1
vlambda=seq(0.1,5,0.1)
lvl=length(vlambda)
mt=matrix(0,nc=length(p),nr=lvl)
for(i in 1:lvl)
        {
        lambda=vlambda[i]
        t=p*sig*sqrt(1/(1-p^2))/lambda
        mt[i,]=t
        }
print(mt)
```

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 450-478.

ISSN 1538 − 9472

Statistical Software Applications and Review:

The Single-Case Data Analysis Package: Analysing Single-Case Experiments with R Software

Isis Bulté
KU Leuven
Belgium

Patrick Onghena
KU Leuven
Belgium

The RcmdrPlugin.SCDA plug-in package is discussed. It integrates three R packages in the R commander interface: SCVA (for Single-Case Visual Analysis), SCRT (for Single-Case Randomization Tests), and SCMA (for Single-Case Meta-Analysis). This way the plug-in package covers three important steps in the analysis of single-case data.

Keywords: Single-case studies, data analysis, software, R package, R commander plugin, GUI

Introduction

To investigate research questions in educational, behavioral, and medical science, single-case experiments are very well suited. To bring these experiments to the attention of (applied) researchers a software package is suggested to analyze data resulting from single-case experiments.

Single-case designs are increasingly popular in educational, behavioral, and medical research (Hammond & Gast, 2010; Matson, Turygin, Beighley, & Matson, 2012; Shadish & Sullivan, 2011; Swaminathan & Rogers, 2007). Bliss, Skinner, Hatau, and Carroll (2008), for example, classified all articles published in four school psychology journals (*School Psychology Quarterly, School Psychology Review, Journal of School Psychology,* and *Psychology in the Schools*) between 2000 and 2005 and found that, with the exception of 2004,

Dr. Bulté is a Researcher in the Faculty of Psychology and Educational Sciences, Methodology of Educational Sciences. Email her at: isis.bulte@ppw.kuleuven.be. Dr. Onghena is a full professor in the Faculty of Psychology and Educational Sciences, Methodology of Educational Sciences. Email him at patrick.onghena@ppw.kuleuven.be.

single-subject designs were more prevalent than group designs. In this time frame, 55% of the causal-experimental studies in these journals utilized single-subject designs. This may seem odd given the strong emphasis on large *N* designs in most statistical and methodological courses and handbooks for the educational, behavioral, and medical sciences.

A possible explanation for the findings of Bliss et al. (2008) lies in the fact that single-case designs can provide a viable alternative or supplement to group designs to answer causal questions. This can be said about educational, behavioral, and medical research in general, but the published applications of single-case designs are certainly not that predominantly present in all subareas. In many subareas of educational, behavioral, and medical research single-case designs however there is a huge potential for single-case designs to complement the standard group designs because of the necessity of pilot data in early stages of larger group studies, because of the relevance of research concerning rare types of participants (e.g., patients with a very specific neuropsychological disorder due to brain injury), or the examination of idiographic questions like 'does this intervention (e.g., restructuring) work for this particular organization?'; and, of course, when research funds are scarce and it is not possible to obtain enough participants for large-scale group studies (Barlow et al., 2009; Edgington & Onghena, 2007; Franklin, Allison, & Gorman, 1997; Kazdin, 2011). Because of the close link of single-case evaluations to individual care, single-case designs are also ideally suited to bridge the scientist-practitioner gap (Barlow, Hayes, & Nelson, 1984; Bliss et al., 2008).

Unfortunately, most of the commonly used statistical software packages, like SPSS and SAS, do not present readily available procedures or options for designing single-case experiments and analyzing single-case data. To fill this gap, an R package for designing single-case experiments and analyzing single-case data is presented.

R was chosen as the computational environment, because it is open source software, running on a variety of UNIX platforms, as well as on Windows and MacOS (Hornik, 2012). R has excellent graphical possibilities, but is also a very powerful and flexible statistical environment, which facilitates the combination of visual and statistical data analysis (Kelley, 2007). However, because of the use of a standard command line interface, R is not very user-friendly. Especially for practitioners who never engaged in any basic programming, the threshold to start working with R can be too high.

Fortunately, Fox (2005) created a window-based graphical user interface (GUI) to R, called the "R Commander" with menus and dialog boxes (very

similar to, e.g., SPSS), which are far more familiar to most people (see Figure 1). Working with the R commander is hands-down: by selecting the menus, submenus open which lead to dialog boxes. Each dialog box contains a 'help' button, which refers the user to a help page with more information. By making selections in the dialog boxes, R commands are generated and executed.

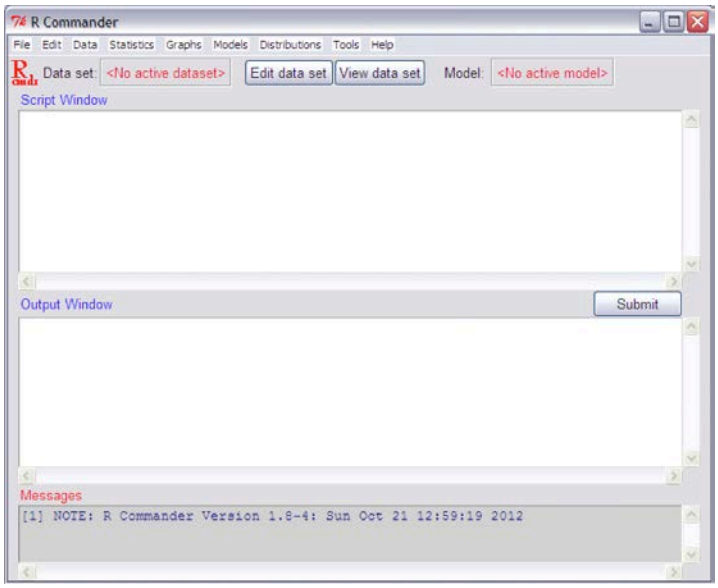


Figure 1. The R commander interface upon starting up.

An advantage of the R commander is that these R commands are not just executed 'behind the scene'. They also appear in the *script window*, so R users can always see the code and adapt it if necessary. This visibility of commands can also be a useful first step in learning R, by gently getting acquainted with the R language. In this script window, also other commands can be typed, or previous commands can be rerun. In the *output window*, the given command appears together with the output, and in the *messages window*, error messages, warnings, and notes appear (Fox, 2005).

An additional advantage of the R commander is that users can add their own menus and dialog boxes. This extensibility became even more practical with the possibility of writing plug-in packages (Fox, 2007). Besides the already mentioned advantages of R and the R Commander, a huge advantage is that all functionalities that are already available in the standard R Commander and in

other plug-in packages can be used. Instead of developing a stand-alone GUI, therefore a plug-in for the R Commander was created. The SCDA (Single Case Data Analysis) plug-in created is a GUI for three R packages presented: SCVA (Single Case Visual Analysis; Bulté & Onghena, 2012), SCRT (Single Case Randomization Tests; Bulté & Onghena, 2008, 2009) and SCMA (Single Case Meta-Analysis; Bulté et al., submitted).

The RcmdrPlugin.SCDA Package: An Illustration

When planning an experiment, first the study design should be carefully chosen based on, among other things, the research question. Then data can be collected according to the selected design. Single-case data analysis, just like any other analysis of empirical data, best starts with a visual exploration of the data. Unless the visualization is that convincing that the effect of the intervention is very obvious (e.g., a dramatic shift in level without any variability or trend), statistical data analysis might be a useful supplementary technique. And often it is not only useful to know whether an intervention had a statistically significant effect, but also what the size of the effect was. The SCDA plug-in for the R commander provides R functions for each of those steps of research design and data analysis. It adds one menu item to the R commander ("SCDA"), which contains three submenus (see Figure 2). Each of the submenus leads to several menu items.

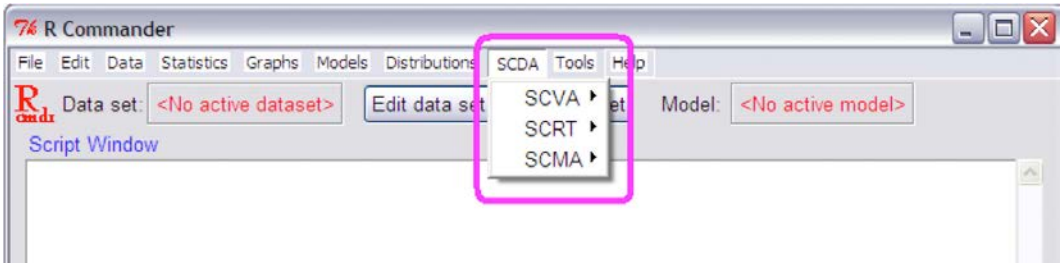


Figure 2. The R commander with the SCDA plugin.

1) SCVA (Single-Case Visual Analysis)
   - Graphical display
   - Plot measure of central tendency
   - Plot estimate of variability

- Plot estimate of trend

2) SCRT (Single-Case Randomization Tests)
   - Design your experiment
     - Number of possible assignments
     - Display all possible assignments
     - Choose 1 possible assignment
   - Analyze your data
     - Observed test statistic
     - Randomization distribution
     - P-value

3) SCMA (Single-Case Meta-Analysis)
   - Calculate effect size
   - Combine *p*-values

What follows will illustrate the functionality of the package with an example from ter Kuile et al. (2009). A more detailed manual can be found in the Appendix.

Ter Kuile et al. (2009) investigated the effectiveness of therapist-aided exposure for lifelong vaginismus. Lifelong (or primary) vaginismus occurs when a woman has never been able to have sexual penetrative intercourse during her whole life. From a cognitive-behavioral perspective, fear and avoidance behavior are connected to vaginismus. Therefore it was hypothesized that exposure to the feared stimuli (i.e., penetration) would increase the ability to have sexual intercourse. The exposure therapy was aided by a trained female therapist and consisted of vaginal penetration exercises (performed by the woman herself) at the hospital, together with several specific homework assignments in which the partner was involved.

Research Design and Data Collection

Ter Kuile et al. (2009) used a replicated single-case AB-phase design to test whether the therapy would lead to an increase in sexual intercourse. In an AB-phase design all baseline measurements (A) precede all treatment measurements

(B). The idea behind this design is to be able to attribute a change in intercourse behavior after the exposure onset to the therapy. To control for time-related confounding variables, randomization was incorporated in the study design: the random aspect was the start of the exposure therapy. This random determination of therapy onset is of course not unlimited. To decide whether the therapy had any effect on the ability to have intercourse, baseline as well as treatment observations are needed. Therefore, in this illustration a constraint of a minimum of one week of diary recordings in each phase is guarded. With a total of 24 weeks of data recordings, this leads to 23 possible start points for the therapy (after week 1, after week 2, …, after week 23) (Figure 3: SCRT -> Design your experiment -> Number of possible assignments).

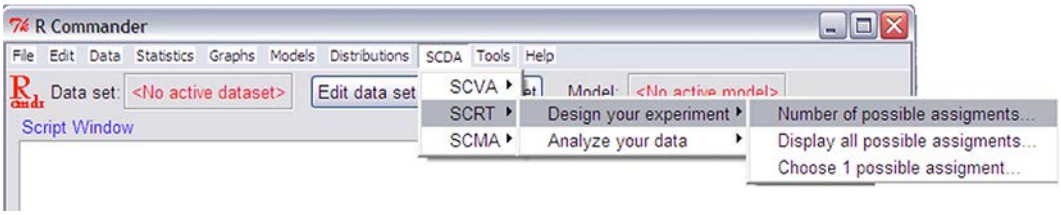


Figure 3. The RcmdrPlugin.SCDA menu, with the SCRT (Single-Case Randomization Tests) 'design your experiment' submenu.

By replicating this experiment over several participants, the strength of the findings is increased and more general statements can be made. Ten patients in sexology clinics, who suffered from primary vaginismus, participated in the study. They kept a daily diary in which they noted (amongst other things) whether they were able to have sexual intercourse with their partner that day. To analyze these data, they should be put into the R commander. This can be accomplished by reading in a created text file with the observations or by entering the observations directly as an active data set (see Appendix).

First the focus is on the diary data of the first patient (P1). She started recording her sexual intercourse attempts from the moment she was referred to the outpatient sexology clinic. The start of the exposure therapy was randomly determined after seven weeks of 'baseline' and she continued filling in the diary for seventeen more weeks.

SCVA (Single-Case Visual Analysis)

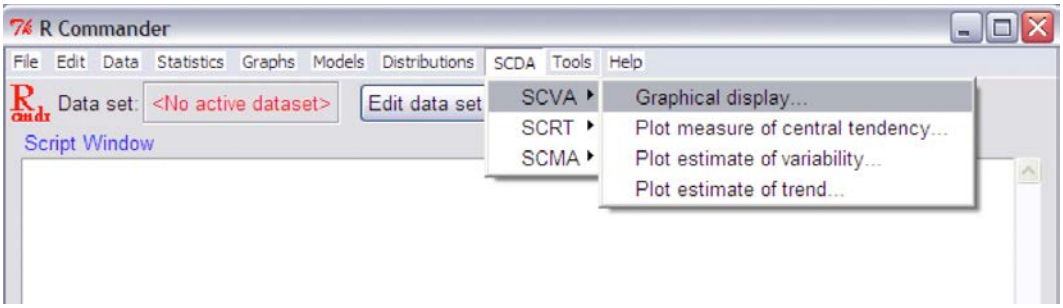


Figure 4. The RcmdrPlugin.SCDA menu, with the SCVA (Single-Case Visual Analysis) submenu.

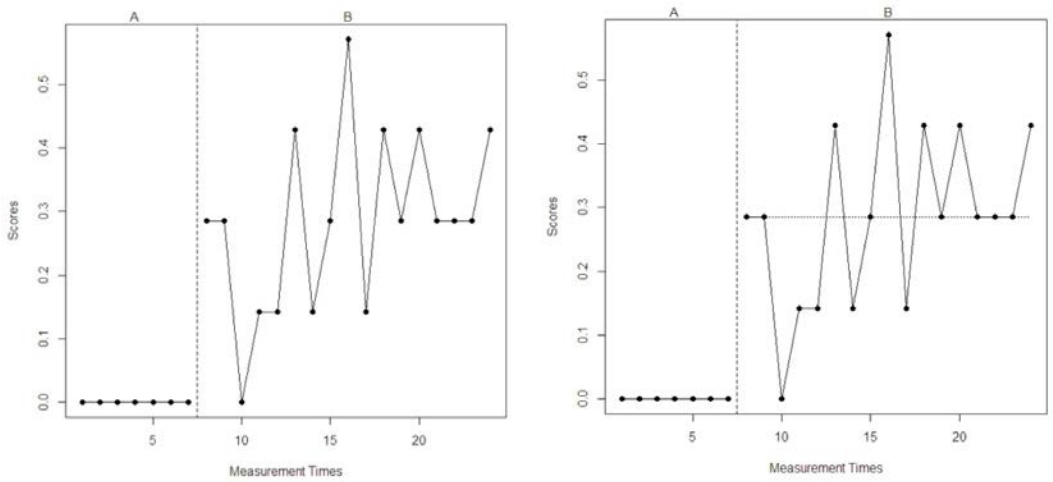


Figure 5. Graphical representation of the diary data of P1, averaged per week (SCDA -> SCVA -> Graphical display; see Figure 4). On the right, the shift in mean level is visualized (SCDA -> SCVA -> Plot measure of central tendency; see Figure 4).

A first impression can be obtained by making a graphical representation of her data (SCDA -> SCVA -> Graphical display; see Figure 4). The left panel of Figure 5 shows the average number of successful intercourse attempts per week for P1. The dotted vertical line indicates the start of the intervention phase. The introduction of the exposure therapy clearly made a difference: before therapy this woman had never experienced sexual intercourse with penetration. The shift in

mean level between the phases is illustrated in the right panel of Figure 5, where the phase means are plotted as a horizontal reference line on the raw data.

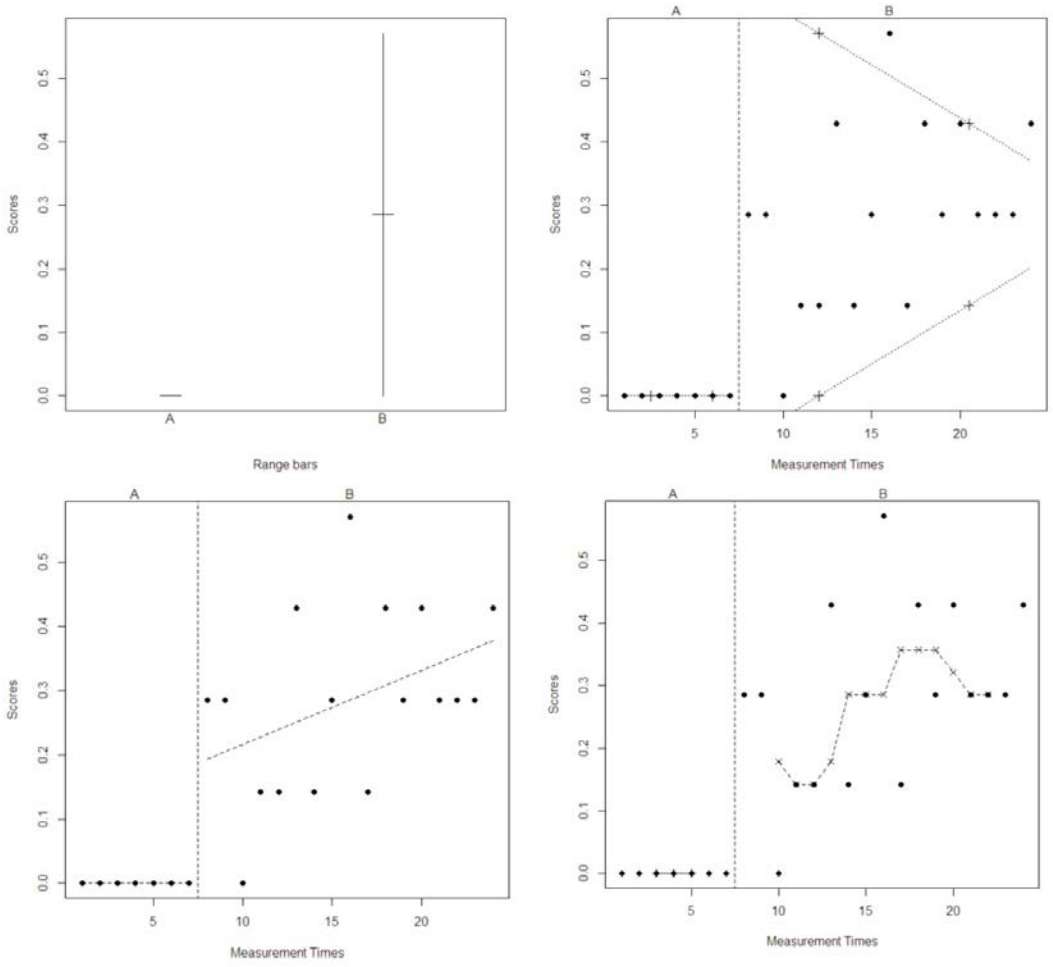


Figure 6. The upper panels illustrate the variation in the data (SCDA -> SCVA -> Plot estimate of variability; see Figure 4): range bars on the left and trended range lines on the right. Possible trends in the data are made visible in the lower panels (SCDA -> SCVA -> Plot estimate of trend; see Figure 4). On the left a linear function is drawn on the raw data by means of ordinary least-squares regression and on the right a nonlinear trend is visualized by displaying running medians of batch size four averaged by pairs.

The range bars in the upper left panel of Figure 6 illustrate the lack of variability in the baseline phase. In the treatment phase variation is shown by the vertical line connecting the minimum and the maximum value, with a small

horizontal bar marking the phase mean. The trended range lines depicted in the upper right panel of Figure 6 show that the variability decreases over time. There is also an upward trend in level noticeable from the linear function in the lower left panel of Figure 6. That this trend is not entirely linear can be seen from the nonlinear smoothed curve produced by calculating running medians of batch size four and averaging each successive pair.

SCRT (Single-Case Randomization Tests)

Visual analysis is an important first step when evaluating intervention effects. In addition, several statistical tests might be conducted to evaluate the statistical significance of the intervention effects. The SCDA GUI includes functions to conduct randomization tests: permutation tests based on random assignment, to test a null hypothesis about treatment effects in a randomized experiment (Onghena & Edgington, 2005). The alternative hypothesis in the illustration presented was that the exposure therapy will lead to an increase in successful sexual intercourse attempts. In other words, the mean of the treatment observations is expected to be higher than the mean of the baseline observations. Therefore, the difference between those means was chosen as test statistic.

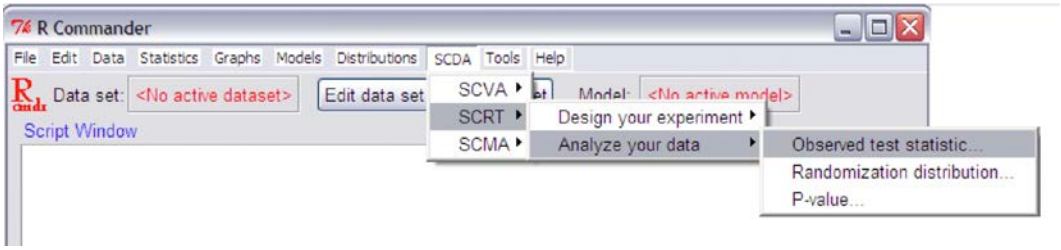


Figure 7. The RcmdrPlugin.SCDA menu, with the SCRT (Single-Case Randomization Tests) 'analyze your data' submenu.

Visual analysis of the data already indicated that the exposure therapy had a positive effect on the intercourse frequency of P1. This effect is also statistically significant, shown by the randomization test's *p*-value of .04 (Figure 7: SCRT -> Analyze your data -> P-value). More information on randomization tests can be found, for example, in Bulté and Onghena (2008).

SCMA (Single-Case Meta-Analysis)

The combination of visual analysis and statistical significance testing, however, does not tell the whole story. It is often also useful to know what the size of the effect is. How large an effect is, is expressed by means of effect size measures (Figure 8: SCDA -> SCMA -> Calculate effect size). The pooled standardized mean difference, which uses the pooled standard deviation of both phases, for P1 equals 2.35. The percentage of data in the treatment phase that is higher than the median of the baseline phase (=”PEM”) is 94%.

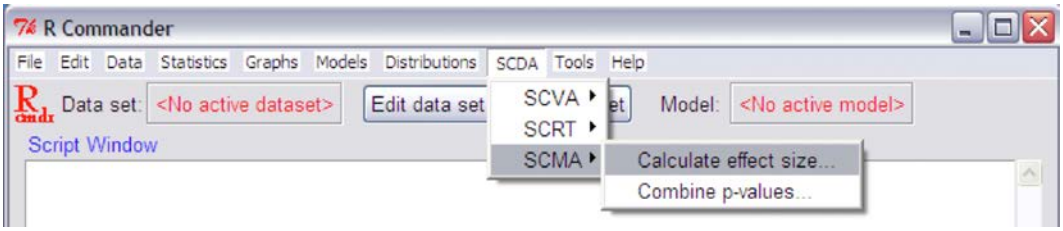


Figure 8. The RcmdrPlugin.SCDA menu, with the SCMA (Single-Case Meta-Analysis) submenus.

These effect sizes are not only used to show the magnitude of an effect, but they also help comparing results over different studies and they can be combined in meta-analyses to come to one general estimate of the effect size. Another meta-analytical procedure is statistically combining the *p*-values of several sequentially replicated single-case experiments, to determine whether a general significant result is obtained (Figure 8: SCDA -> SCMA -> Combine P-values). Ten women with primary vaginismus participated in the study of ter Kuile et al. (2009). When combining their results with the additive method for combining independent *p*-values, a general significant result of $p = .0016$ was obtained.

Conclusion

The use of the SCDA plugin package for the R commander was illustrated. A more elaborated overview of all functionalities and an explanation of the different parts in the dialogue windows is given in the Appendix.

The presented software package contains one technique for statistically analyzing the data resulting from single-case experiments. These randomization tests have several advantages over other techniques suggested in the literature.

Most importantly, the presence of serial dependency in the data will not invalidate the result of a randomization test. This is of considerable interest in the context of single-case experiments, because these data tend to have autocorrelated residuals, which can seriously bias the results of for example *t*-tests, by inflating the Type I error rate. Another advantage is that they are free from the assumption of random sampling. This random sampling assumption, on which the probability tables of parametric tests are built, is usually an unrealistic ideal and most experiments do not use randomly sampled subjects. Also, unlike parametric *t*-test, randomization tests are not based on assumptions about the homogeneity of variances. But they do have one requirement: the incorporation of random assignment not only enhances the internal validity of the study, as indicated before, but it also justifies the application of a randomization test that is based on the random assignment used in the experimental design of the study. This way it is possible to infer causal relations between the treatment and the observed changes in behavior. Because randomization tests acquire validity by mirroring the random assignment schedule used in the study, their extreme versatility permits researchers to make valid tests by devising a test that is suitable for the particular design used. They can be used with all sorts of data (continuous, discrete, ranks,...) and all kinds of data patterns (trends, outliers, skewed distributions, zero variance in baseline,...) (see e.g., Edgington, 1973, 1980; Edgington & Onghena, 2007; Gorman & Allison, 1997; Kazdin, 2011; Ludbrook, 1994; Onghena & Edgington, 2005; Dugard, File & Todman, 2012).

An alternative to these nonparametric tests is time series analysis, which investigates whether there is a significant change in level and/or slope between the phases. This technique is also suitable for the analysis of data when serial dependency is present. It, however, requires many data points to determine the existence and the pattern of autocorrelation and to identify the model correctly, which could cause problems for small-*n* experiments where the phases are usually rather short. Another difficulty is the complexity of the mathematical theories on which it is based (Box, Jenkins & Reinsel, 1994; Gorman & Allison, 1997; Kazdin, 2011). Another alternative is hierarchical linear modeling (HLM). Van den Noortgate and Onghena (2003a, 2003b) suggested using such a model for calculating effect sizes and for combining effect sizes of single-case data or combining the raw data of several studies. By modeling the hierarchical structure, the possible dependence of the scores is taken into account. Also study or case characteristics can be included as covariates to explain possible variation. For time series analysis a plug-in for the R commander already exists

(RcmdrPlugin.epack); this is not the case for hierarchical linear modeling, but R packages do exist (e.g., ‘nlme’).

As a useful extension of the RcmdrPlugin.SCDA package, these techniques could be included. Not only the HLM suggestion of Van den Noortgate and Onghena (2003a, 2003b), but also other methods for combining effect sizes could be adopted in the package. The simplest method is probably just taking the (weighted) average of all effect sizes to obtain one overall measure that reflects the general effect of the intervention over the different studies. This could for example be done with the effect size estimates already included in the package. Additionally, more effect size measures could be added, such as the percentage of all non-overlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007). Other possibilities for combining effect sizes include the three approaches presented by Busk and Serlin (1992), which differ in the assumptions made about the distribution and the variability of the data. Also the inclusion of more methods for combining *p*-values could be interesting: Stouffer’s method, in which the *Z*s associated with the *p*-values are added and divided by the square root of the number of studies, after which the resulting *Z*-value is converted to an overall *p*; Mosteller and Bush’s modification, that computes a *t*-test on the obtained *Z* value; Winer’s suggestion of adding *t*-values and dividing the sum by the square root of the degrees of freedom (Rosenthal, 1978); and the iterative procedure for combining *p*-values, by applying more than one combining function to the same partial tests and then combining the resulting second order *p*-values into a third order of combination by means of a combining function, until the final *p*-value becomes reasonably invariant (Pesarin, 2001; Pesarin & Salmaso, 2010).

References

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The scientist practitioner: Research and accountability in clinical and educational settings.* New York: Pergamon.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Allyn & Bacon.

Bliss, S. L., Skinner, C. H., Hautau, B., & Carroll, E. E. (2008). Articles published in four school psychology journals from 2000 to 2005: An analysis of experimental/intervention research. *Psychology in the Schools, 45*, 483-498.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467-478.

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41*, 477-485.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology, 8*, 104-114.

Bulté, I., Van den Noortgate, W., Heyvaert, M., & Onghena, P. (submitted). Meta-analysis of single-case studies: The SCMA package. Manuscript submitted for publication.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum.

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests* (2nd ed.). New York: Routledge Taylor & Francis Group.

Edgington, E. S. (1973). The random-sampling assumption in "Comment on component-randomization tests." *Psychological Bulletin, 80*, 84-85.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics*, **5**, 235-251.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). London: Chapman & Hall/CRC.

Fox, J. (2005). The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software, 14.*

Fox, J. (2007). Extending the R commander by "plug-in" packages. *R News*, 7(3), 46-52. URL http://CRAN.R-project.org/doc/Rnews/.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Design and analysis of single-case research.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gorman, B. S. & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: Erlbaum.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single-subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities, 45,* 187-202.

Hornik, K. (2012). *The R FAQ: Frequently asked questions on R.* Retrieved February 17, 2013 from cran.r-project.org/doc/FAQ/R-FAQ.html.

Huber, P., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press, 2010.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39*, 979-984.

Ludbrook, J. (1994). Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clinical & Experimental Pharmacology & Physiology, 21,* 673-686.

Matson, J. L., Turygin, N. C., Beighley, J., & Matson, M. L. (2012). Status of single-case research designs for evidence-based practice. *Research in Autism Spectrum Disorders, 6,* 931-938.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21,* 56-68.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40,* 194-204.

Pesarin, F. (2001). *Multivariate permutation tests. With applications in biostatistics.* New York: John Wiley & Sons.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software.* Chichester, UK: Wiley.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971-980.

Swaminathan, H., & Rogers, H. J. (2007). Statistical reform in school psychology research: A synthesis. *Psychology in the Schools, 44,* 543-549.

ter Kuile, M. M., Bulté, I., Weijenborg, P. T. M., Beekman, A., Melles, R., & Onghena, P. (2009). Therapist-aided exposure for women with lifelong vaginismus: A replicated single-case design. *Journal of Consulting and Clinical Psychology, 77,* 149-159.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*, 1-10.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Elsevier Academic Press.

Appendix: Getting started with the RcmdrPlugin.SCDA

Installation

To make the SCDA GUI work, R should be downloaded and installed. This can be done at no cost from the CRAN (Comprehensive R Archive Network; cran.r-project.org) website. Hornik (2012) gives a detailed explanation of how to do this for Windows, Macintosh and UNIX (the R package presented here is, however, only tested on Windows). Once R is running, the GUI can be installed from within the R console (Packages -> Install packages). After choosing a CRAN mirror nearby, the 'RcmdrPlugin.SCDA' package should be selected from the list. Note that the downloading process can take a while because several supporting packages are downloaded automatically as well. Then the RcmdrPlugin.SCDA package can be loaded into R by selecting 'Packages -> Load package' (or by typing the command *library('RcmdrPlugin.SCDA')* into the R console). Only this last step needs to be repeated when using the GUI. Note that under Windows, the R commander functions best with the single-document interface (SDI: Edit -> GUI preferences -> Single or multiple windows: SDI). By loading the RcmdrPlugin.SCDA package, the R commander opens with 'SCDA' as an additional menu button (Figure A1).

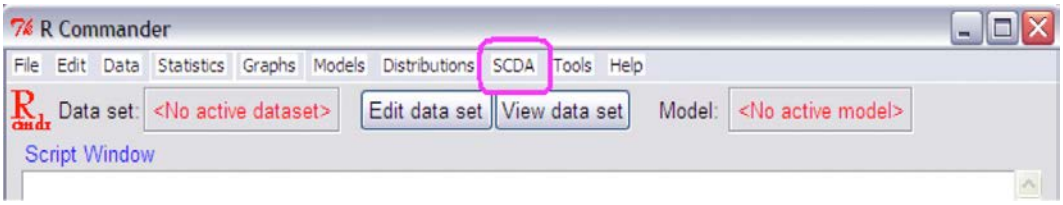


Figure A1. By loading the RcmdrPlugin.SCDA package, the "SCDA" menu button is added to the R commander interface.

Data Input

Most of the functions in the SCDA GUI need data input, which can be taken from a .txt file that has been created in advance in a text editor (e.g., EditPad or NotePad) or in Excel (save the file as 'Text(Tab delimited)'). Text files containing the raw data should consist of two tab-separated columns for single-

case phase and alternation designs: the first with the condition labels ("A" and "B" when there are two conditions/phases, and "A1", "B1", "A2" and "B2" for three or four phases) and the second with the observations. For multiple-baseline designs they should consist of these two columns for each unit. This way, each row represents one measurement occasion. Text files containing *p*-values for the 'combine' function should consist of one column of *p*-values. In text files containing the possible start points for multiple baseline designs, each row should contain all possibilities for one unit, separated by a tab. It is important not to label the rows or columns.

Another way of data input is by using the active data set of the R commander. Creating an active data set has the advantage that also other functions than the SCDA-functions (e.g., built-in statistical functions of the R commander or from other plug-ins) can be applied to the data. This way one can fully benefit from the functionality of the R commander and its plug-ins. There are several ways to construct this active data set: a text file created as described above can be read into the R commander via the Data menu (Figure A2: Data -> Import data -> from text file, clipboard, or URL -> uncheck the box 'Variable names in file' if necessary -> OK -> select the text file within the 'Open file' dialog box), or data can be entered directly via 'Data -> New data set'. The active data set can be consulted at any time by clicking the 'View data set' button.

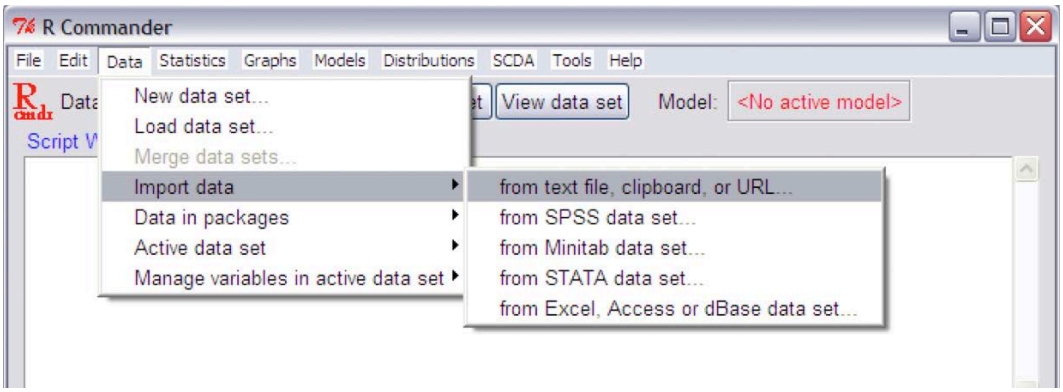


Figure A2. Constructing the active data set by reading a text file into the R commander.

Functions

The SCDA menu contains three submenus with several menu items. Each menu item opens a dialogue box.

SCVA (Single-Case Visual Analysis)

Graphical display The observed single-case data are plotted. As can be seen in Figure A3, two selections should be made in the dialogue box. In section A, the *design type* used has been indicated. The options are:

Phase Designs
Comparisons are made within a time series and the subject's performance is evaluated over time across baseline (A) and intervention (B) phases.

AB Phase Design	*ABA Phase Design*:	*ABAB Phase Design*
All baseline measurements precede all treatment measurements.	Withdrawal or reversal design in which the treatment is administered between two baseline phases.	An extra intervention phase is added.

Alternation Design
The basic strategy is the rapid alternation of two or more conditions within a single subject.

Completely Randomized Design	*Alternating Treatments Design*	*Randomized Block Design*
The random assignment procedure mirrors the one used when randomly assigning subjects to different groups for an independent samples *t* test (no restrictions).	The temporal clustering of treatments is prevented by ensuring that the randomization does not permit more than a specified number of successive time blocks with the same condition.	Adjacent treatment times are grouped together in blocks and the conditions are assigned randomly within each block.

Multiple Baseline Design
This is an extension of the basic AB phase designs, in which several of those AB designs are implemented simultaneously to different persons, behaviors, or settings. A characteristic feature is that the intervention is introduced in a staggered way to the different units.

In section B, the *data file* in which the data can be found should be selected. This could either be the active data set or a previously created text file. In the latter case, the file can be chosen with the 'Select File' button.

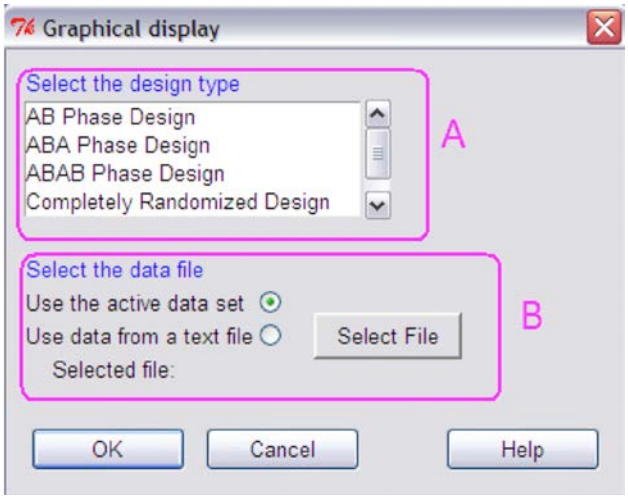


Figure A3. The dialogue box for the 'graphical display' menu.

Plot measure of central tendency a measure of central tendency is plotted as a horizontal reference line superimposed on the raw time series data.

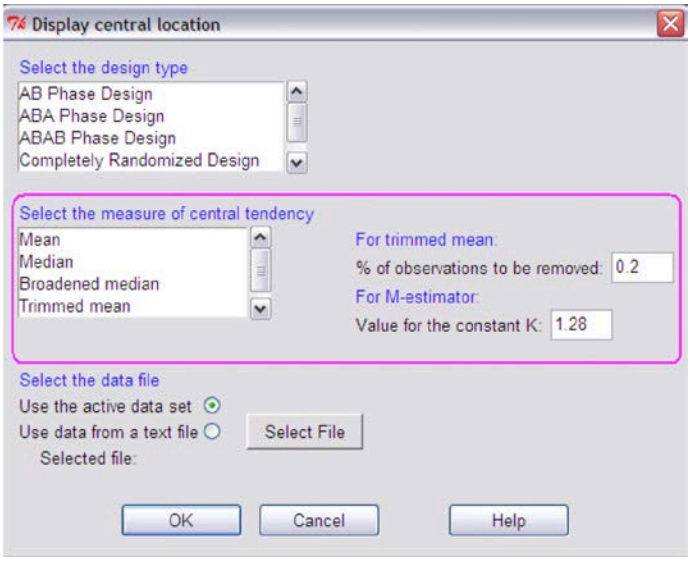


Figure A4. The dialogue box for the 'Plot measure of central tendency' menu.

In the marked section of the dialogue box (see Figure A4), the *measure of central tendency* should be selected. There are four built-in possibilities:

Mean
The arithmetic mean (the average) of the data. This measure is rather sensitive to outliers.

Median
The middle value of the data (for an even number of data points, the median equals the average of the two central data points). This measure is more robust than the mean, but it only takes into account the central data points, while disregarding other numerical information.

Broadened median
Calculated based on the three, four, five or six middle values of the data set (depending on the total number of data points). This measure is more robust than the mean, and sensitive to a larger proportion of the data than the median.

Trimmed mean
Calculates the mean after discarding the extreme observations. The percentage of observations that has to be removed from each end of the distribution can be indicated in the upper right box (Figure A4). This can be any value from 0 (= regular arithmetic mean) to 0.5. The default value is 20 percent. The trimmed mean is more robust than the mean and less affected by observations in the centre of the distribution than the (broadened) median.

M-estimator
Huber's M-estimator of location first evaluates each observation to determine if it is actually an outlier compared to the rest of the data and then gives less weight to those outlying values. For this evaluation a constant K needs to be specified that can have any value between 0 and ∞. Usually a percentile of the standard normal distribution is chosen. Wilcox (2005) suggests using $K = 1.28$, which corresponds to the 90th percentile of the standard normal distribution and covers 80 percent of the underlying distribution. By determining the sensitivity of the estimator, one can balance between robustness and efficiency (see e.g., Huber & Ronchetti, 2009).

Plot estimate of variability Information about variability in the data is displayed by one of three methods.

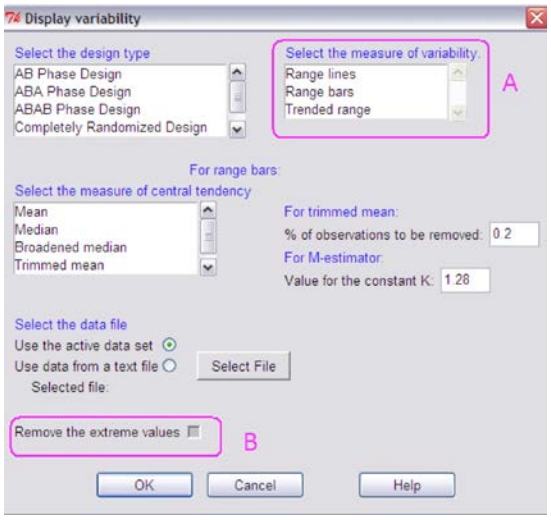


Figure A5. The dialogue box for the 'Plot estimate of variability' menu.

In box A (Figure A5) the *measure of variability* should be selected. The three options are:

Range lines
Range lines are a pair of lines parallel to the X-axis, passing through the lowest and highest values for each phase, and superimposed on the raw data.

Range bar graphs
Range bar graphs consist of a vertical line for each phase, connecting the minimum and the maximum value, with a small horizontal bar crossing this line to display a measure of central location ((trimmed) mean, (broadened) median, or M-estimator). This estimate of central tendency should be selected in the middle part of the dialogue box (Figure A5).

Trended ranges
Trended ranges display changes in variability within phases by two lines, one connecting the minimum values of the phase halves and one connecting the maxima.

For all these methods the *influence of outliers* may be lessened by using a trimmed range, in which only a sample of the data set is used. This can be selected in box B (Figure A5): default the whole dataset is used, but by checking the box the 10 to 20 percent extreme values are removed from each phase.

Plot estimate of trend This function visualizes systematic shifts in central location over time using several methods. The method of *trend visualization* should be selected (Figure A6):

Vertical line plot
A *vertical line plot* draws the deviations from each data point to a measure of central tendency against time. The measure of central tendency should be selected in the middle part of the dialogue box (Figure A6).

Trend lines

Trend lines superimpose a linear function on the raw data, which shows if there is an increase or a decrease in the observed behavior over time.

Least squares regression
Minimizes the squared vertical distances between the regression line and the data points.

Split-middle method
Connects the crossings of the median dependent variable value and the middle time value of both phase halves

Resistant trend line fitting
Comparable to the split-middle method, but here the phases are divided into three sections instead of two. This makes this method more suited for longer time series.

Running medians

The presence of a nonlinear trend can be displayed with *running medians*, with which the time series is smoothed by dividing it into successive segments of a given size and calculating the median for each segment (Tukey, 1977). Three sizes of segments are easy to use with time series data: *batch size 3, batch size 5, and batch size 4 averaged by pairs*.

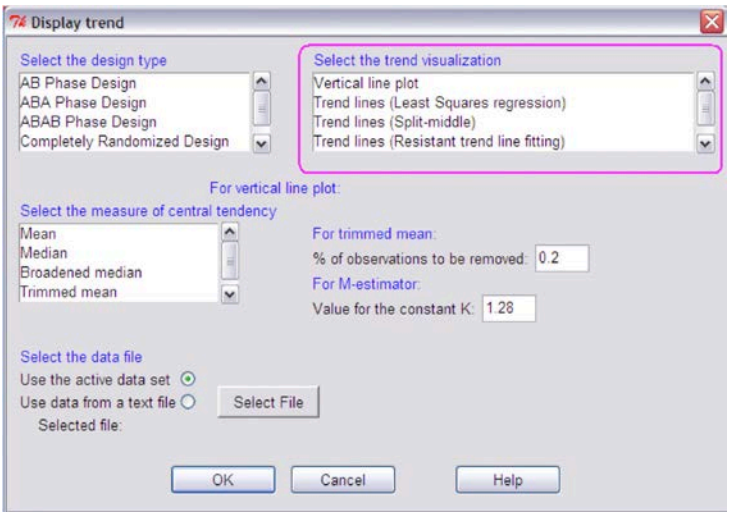


Figure A6. The dialogue box for the 'Plot estimate of trend' menu.

SCRT (Single-Case Randomization Tests)

Design your experiment

Number of possible assignments

The number of assignment possibilities for the specified design is calculated. In box A (Figure A7) the *total number of observations* in the experiments should be indicated. This is not necessary when the design used is a multiple baseline design. In box B a *constraint* on the randomization schedule should be provided. In phase designs the moment of phase change is randomly determined, so the restriction is placed on the minimum number of observations per phase. In alternation designs the observations are randomly assigned to different conditions, so in other words the treatment order is randomly determined. To avoid too long sequences of the same condition, in alternating treatment designs there is a restriction on the maximum number of consecutive administrations of the same condition. For multiple baseline designs, there is one extra selection to be made: in box C the location of the file with the *possible start points* should be selected (see 'Data input').

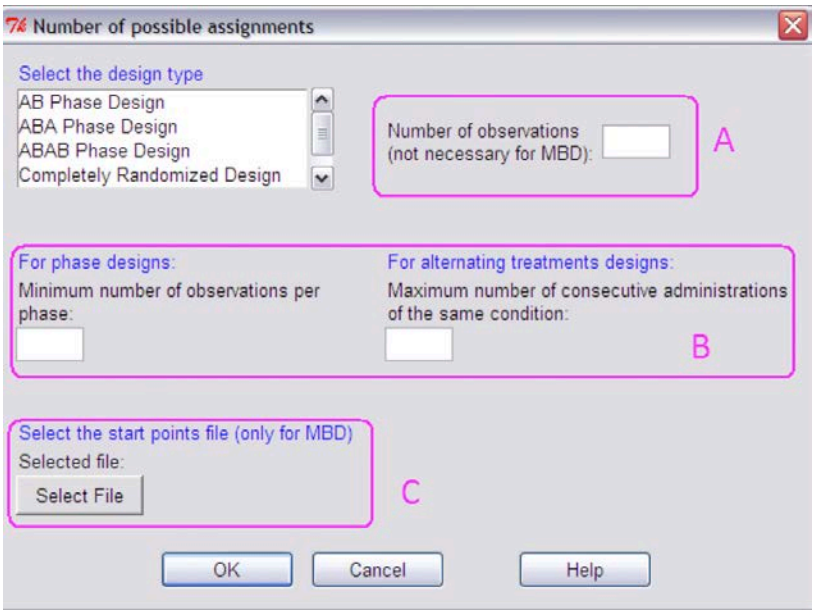


Figure A7. The dialogue box for the 'Number of possible assignments' menu.

Display all possible assignments

All assignment possibilities for the specified design are enumerated. In the dialogue box (Figure A8) can be chosen if the possibilities should only be displayed in the output window of the R commander, or also should be saved to a file. The location to save the assignment possibilities has to be indicated with the 'Select location' button.

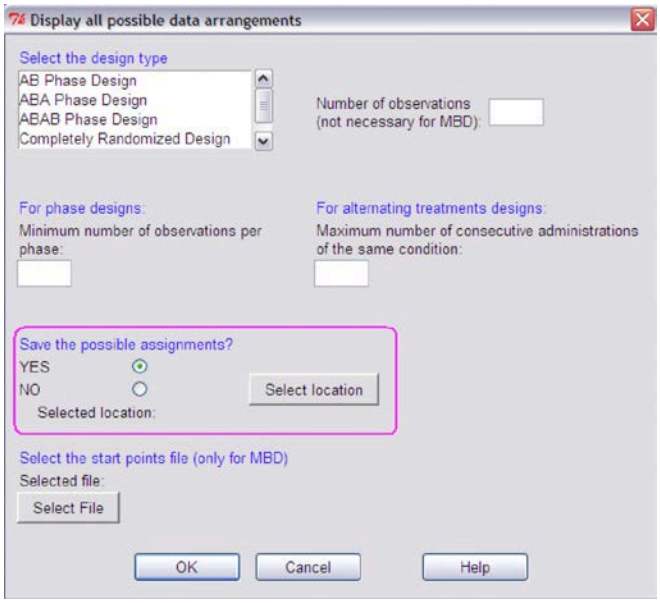


Figure A8. The dialogue box for the 'Display all possible assignments' menu.

Choose 1 possible assignment

One assignment possibility is randomly selected from all theoretical possibilities. The dialogue box is similar to that of the 'Number of possible assignments' function.

Analyze your data

Observed test statistic

The observed test statistic is calculated from the obtained raw data. There are several built-in possibilities as *test statistic* (see Figure A9). For alternation designs, multiple-baseline designs and AB phase designs, there are three

options: *"A-B"*, *"B-A"*, and *"|A-B|"*, which stand for the (absolute value of the) difference between the condition/phase means. For phase designs with more than two phases, six more options are available: *"PA-PB"*, *"PB-PA"*, and *"|PA-PB|* refer to the (absolute value of the) difference between the means of phase means, and *"AA-BB"*, *"BB-AA"*, and *"|AA-BB|"* represent the (absolute value of the) difference between the sums of phase means.

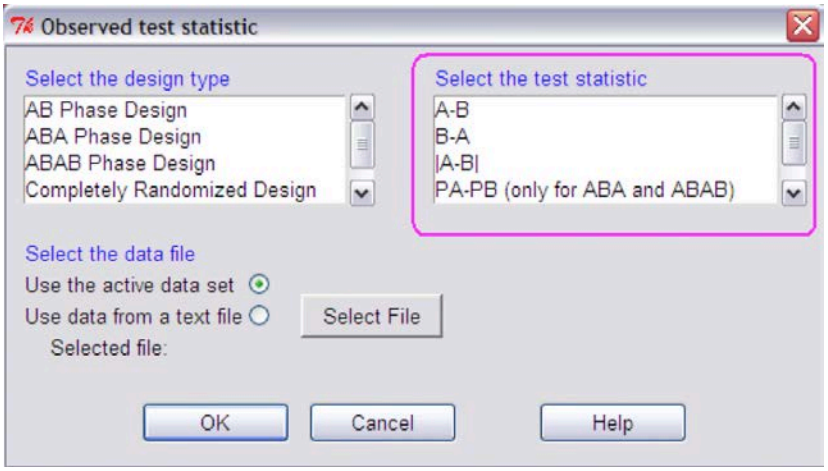


Figure A9. The dialogue box for the 'Observed test statistic' menu.

Randomization distribution

The randomization distribution is generated. One can choose between the exhaustive ('systematic') and the nonexhaustive ('Monte Carlo') randomization distribution (Figure A10 box A). For the exhaustive randomization distribution all assignment possibilities are enumerated, while the nonexhaustive randomization distribution is generated by a random sample of all assignment possibilities. The size of this random sample should be indicated in the box 'number of randomizations'. In box B (Figure A10) can be chosen if the randomization distribution should only be displayed in the output window of the R commander, or also should be saved to a file. The location to save the randomization distribution has to be indicated with the 'Select location' button

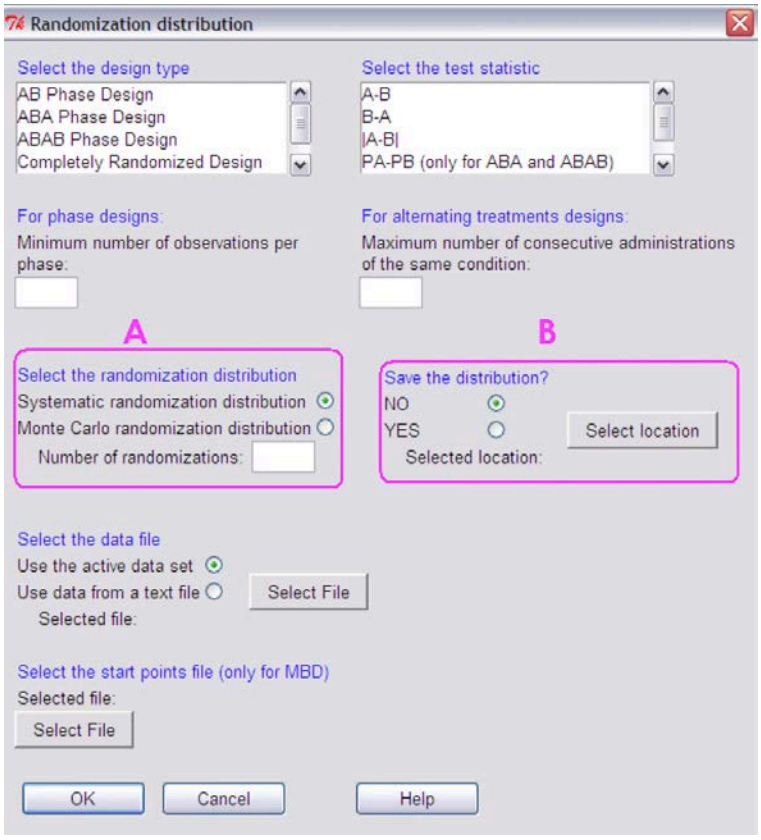


Figure A10. The dialogue box for the 'Randomized distribution' menu.

P-value

The *p*-value corresponding to the observed value of the test statistic is obtained by locating this value in the exhaustive or nonexhaustive randomization distribution. The dialogue box is similar to that of the 'randomization distribution' function.

SCMA (Single-Case Meta-Analysis)

Calculate effect size The specified effect size measure is calculated (Figure A11). Four effect size measures are included in the RcmdrPlugin.SCDA:

Standardized mean difference
The mean of the baseline condition is subtracted from the mean of the treatment condition, and this difference is divided by the baseline standard deviation.

Pooled standardized mean difference
Similar to the standardized mean difference, but with the pooled standard deviation used as denominator

PND (percentage of nonoverlapping data)
The percentage of data points in the treatment phase that exceed the most extreme value in the baseline phase (i.e., the proportion lower than the lowest baseline point for interventions designed to decrease unwanted behavior, and the proportion higher than the highest baseline point for interventions intended to increase desired behavior).

PEM (percentage of data points exceeding the median)
The overlap is calculated as the percentage of data points in the treatment phase that are located above the extended median line of the previous baseline phase (or below this line, if the undesired behavior is expected to decrease).

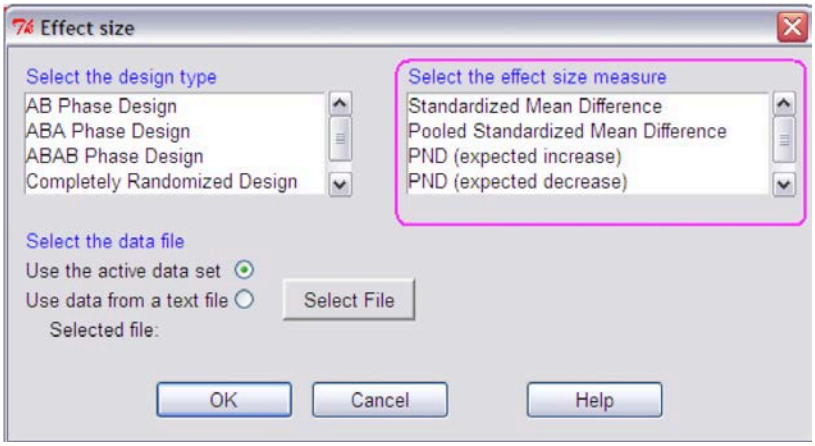


Figure A11. The dialogue box for the 'Calculate effect size' menu.

Combine p-values Nonparametric combination of *p*-values, with the multiplicative approach using Pearson's formula or the additive approach using Edgington's formula (Figure A12 box A). In box B (Figure A12) the location of the file in which the *p*-values to be combined can be found should be selected.

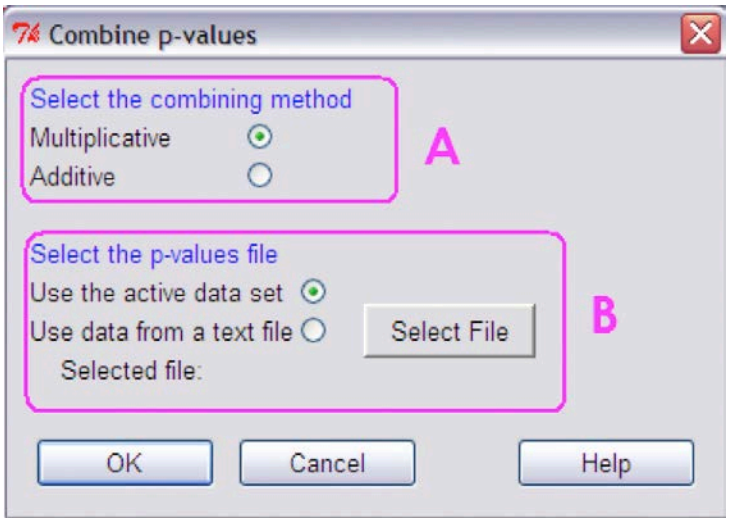


Figure A12. The dialogue box for the 'Combine *p*-values' menu.

בס"ד
Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2

ISSN 1538 − 9472

Instructions for Authors

Authors wishing to submit to *JMASM* may do so using the submission form at the journal's website, http://digitalcommons.wayne.edu/jmasm. Three areas are appropriate for *JMASM*:

1. Development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods;
2. Development or study of nonparametric, robust, permutation, exact, and approximate randomization methods; and
3. Applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in *Regular Articles*, *Brief Reports*, and *Emerging Scholars* is externally peer reviewed, with input from the Editorial Board; work appearing in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* is internally reviewed by the Editorial Board.

Please observe the following guidelines when preparing manuscripts:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
3. Manuscripts should be prepared in Microsoft Word (.doc or .docx) only (Wordperfect and .rtf formats may be acceptable − please inquire). Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are NOT acceptable for manuscript submission.
4. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
5. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
6. The submission form requires an Abstract with a 50 word maximum, and a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left justified, indent optional.

ISSN 1538 − 9472

7. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.

8. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use "&" instead of "and" in multiple author listings.

9. Suggestions for style: Instead of "I drew a sample of 40" write "A sample of 40 was selected". Use "although" instead of "while," unless the meaning is "at the same time." Use "because" instead of "since," unless the meaning is "after." Instead of "Smith (1990) notes" write "Smith (1990) noted." Do not strike the spacebar twice after a period.

Journal of
Modern Applied Statistical Methods

ISSN: 1538−9472 http://digitalcommons.wayne.edu/jmasm

PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

To Advertisers

Advertisements are accepted at the discretion of the editor. Send requests for advertising information to ea@jmasm.com.